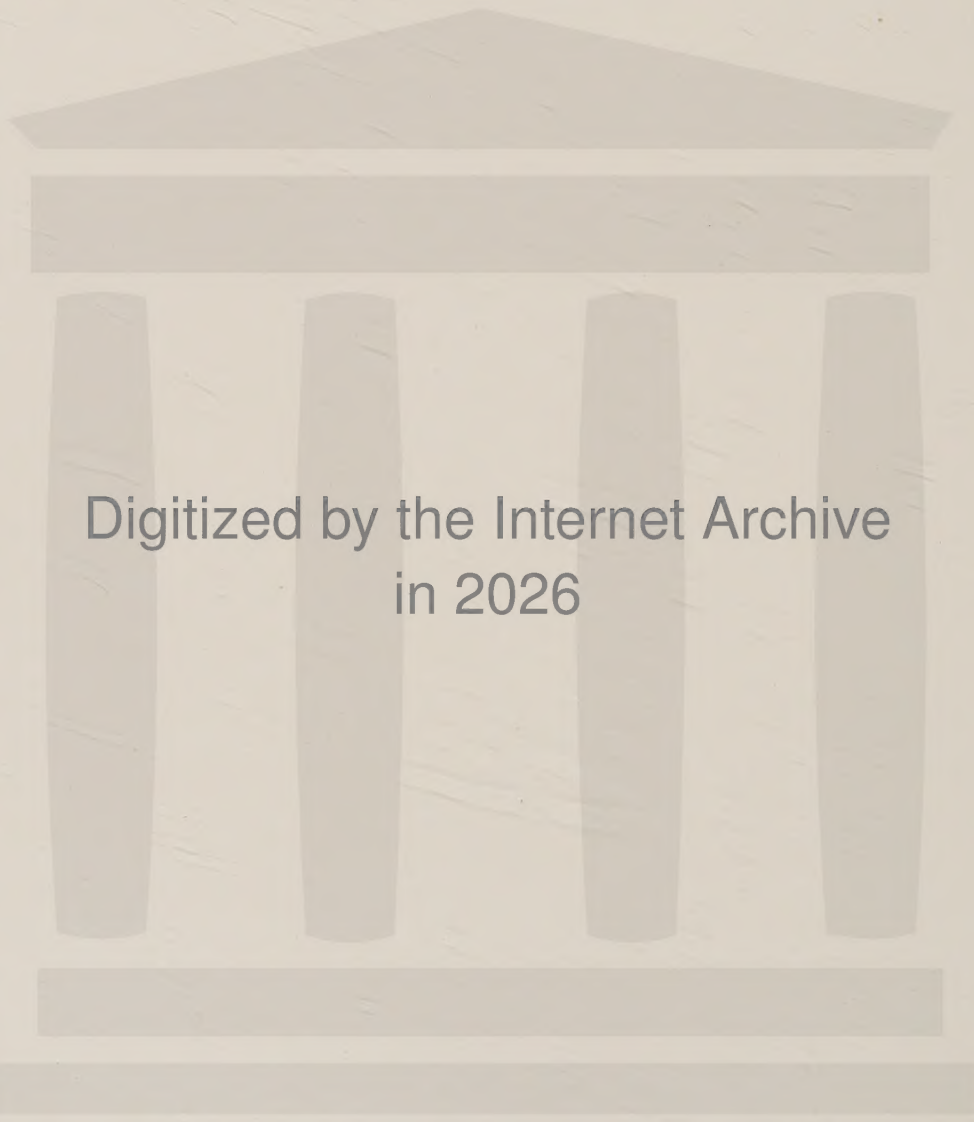




Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41555825_0058

THESES

Présentées

A LA FACULTE DES SCIENCES

DE L'UNIVERSITE DE PARIS.

pour obtenir le titre de Docteur de l'université

(mention Science)

par

ABBAS GHOLI KHADJENOURI

1ère THESE: Contribution à la théorie mathématique de l'échantillonnage: Sur les modifications à apporter à la théorie des sondages pour tenir compte de la nécessité d'arrondir les valeurs théoriques des effectifs de l'échantillon.

2ème THESE: L'orthogonalité et ses applications à la statistique mathématique.

Soutenue le 19 Juin 1956 devant la commission d'examen.

Prof. Georges Darmois

Président

Prof. Daniel Duguet

:
Examineurs

Prof. Robert Fortet

:

Mention ; très honorable.

Monsieur le Professeur Georges DARMOIS a bien voulu encourager nos recherches: qu'il daigne recevoir, ici, l'expression de notre gratitude la plus profonde.

Nous remercions aussi M.P.THIONET à qui nous sommes redevables de nombreux conseils.



A ma chère femme

STEFKA

<u>Table des matières.</u>	pages
<u>Introduction</u>	1
1-Conditions particulières d'une enquête à Téhéran	1
2-Plan proposé pour la future enquête par sondage	3
3-Estimation du total	5
4-Estimation de la variance de Y	7
5-Le problème de nombre optimum des strates	10
6-Application numérique	12
<u>Première partie-Le problème général de l'arrondissement à un entier des valeurs théoriques obtenues pour tel ou tel effectif de sondage</u>	15
1-Introduction	15
2-Etude du problème de base	18
3-Le cas de plusieurs strates	22
4-Cas d'un plan impliquant de nombreux arrondissements ...	26
5-Distribution de ξ	30
6-Distribution de η_1	32
7-Etude comparée de nos diverses méthodes d'arrondissement	34
8-Espérance mathématique de $1/\mu$	45
9-Resume de la première partie	50
<u>Deuxième partie-Les nouvelles formules d'estimateurs et de variances pour divers plans de sondage</u>	51
1-Enoncé du problème général	51
2-Liste des plans étudiés	56
3-Plan II-Sondage à deux degrés avec probabilités égales et tirages exhaustifs	57

4-Plan III-Stratification suivie de deux degrés de son- dage avec probabilités égales et tirages exhaustifs	62
5-Plan IV-Sondage à deux degrés avec stratification des PSU	67
6-Plan V-Sondage de deux PSU avec probabilités inégales et tirages exhaustifs	72
7-Plan VI-Stratification suivie de deux degrés de son- dage, tirage d'une PSU de chaque strate avec les probabi- lités inégales	75
8-Plan VII-Sondage à deux degrés, le premier avec pro- babilités inégales et tirages bernoulliens	80
9-La comparaison des différents plans de sondage au - point de vue de l'augmentation de variance	83
<u>Troisième partie-Répartition optimum d'échantillon</u>	
<u>quand la population a une structure donnée</u>	88
1-Enoncé général du problème	88
2-Fonction de coût.	102
3-Répartition optimum, au point de vue: c constant. ...	106
3.1-Le "problème de Neyman" pour un simple plan de sondage stratifié.	106
3.2-Le "problème de Neyman" pour un simple plan de son- dage stratifié, tirage bernoullien et probabilités inégales dans chaque strate.	112
4-Répartition optimum, au point de vue: V constante... ..	114
4.1-L'équivalent du "problème de Neyman" avec un échan- tillon non plus stratifié mais à deux degrés.	114

4.2-Extension pour K degrés de sondage.	118
4.3-Tirage exhaustif de deux PSU (en tout) et probabilités inégales.	127
4.4-Stratification suivie de deux degrés de sondage, tirage de 1 PSU de chaque strate avec les probabilités inégales.	131
4.5-Sondage à deux degrés n PSU par tirage bernoullien et les probabilités inégales.	134
5-Les problèmes du type 2: choix optimum de r, s, 137	
5.1-Sondage à deux degrés avec probabilités égales et tirages exhaustifs (plan II).	137
5.2-Stratification suivie de deux degrés de sondage par tirages exhaustifs et avec probabilités égales (plan III).	140
5.3-Sondage à deux degrés avec stratification des PSU (plan IV).	143
5.3.1-Premier cas particulier.	144
5.3.2-Deuxième cas particulier.	149
5.4- Deux PSU (en tout), tirages exhaustifs, probabilités inégales (plan V).	154
5.5-Stratification, puis une PSU de chaque strate, avec probabilités inégales (plan VI).	156
5.6-Sondage à deux degrés avec probabilités inégales et tirages bernoulliens (plan VII).	159
6-Resumé de la 3ème partie	161

<u>Quatrième partie-Problèmes de recherche de structure optimum.</u>	1650
1-Enoncé de la forme générale du problème.	165
2-Plan II-Deux degrés de sondage avec les probabilités égales et tirages exhaustifs.	168
3-Plan 3-Sondage stratifié à deux degrés, avec probabilités égales et tirages exhaustifs.	184
4-Plan IV-Sondage à deux degrés, avec stratification des PSU, tirages exhaustifs et probabilités égales.	191
4.1-Premier cas particulier.	192
4.2-Deuxième cas particulier.	194
5-Plan V-Deux PSU(en tout)avec probabilités inégales et tirages exhaustifs.	196
6-Plan VI-Stratification, puis 1 PSU de chaque strate avec probabilités inégales.	198
7-Plan VII-Sondage à deux degrés, avec probabilités inégales et tirages bernoulliens.	202
8-Influence du changement de la structure de la population sur V_2 contribution des erreurs d'arrondissement à la variance.	204
<u>Conclusions.</u>	207
<u>Références.</u>	211

Résumé de la deuxième Thèse	213
L'orthogonalité et ses application à la statistique mathématique	213
Résumé du chapitre 1	214
Résumé du chapitres 2 et 3	216
Résumé du chapitre 4	221
Résumé du chapitre 5	221
Résumé du chapitre 6	224
Résumé du chapitre 7	224
Résumé du chapitre 8	227
Résumé du chapitre 9	228
Résumé du chapitre 10	228
Résumé du chapitre 11	229

I N T R O D U C T I O N

Nous commencerons notre exposé en d^{écr}ivant de façon indépendante le plan de sondage particulier qui est à l'origine de cette thèse. Cet exemple permet de bien comprendre le type général de problèmes qui seront étudiés ensuite dans les 1ère, 2ème, 3ème et 4ème Parties de notre thèse.

Le point de départ de notre thèse est donc un plan de sondage établi en 1955 et destiné à la zone urbaine de Téhéran.

0

0 0

1. Conditions particulières d'une enquête à Téhéran.-

Pendant l'été 1954, nous avons procédé au préalable à une enquête pilote auprès de certaines familles de Téhéran (Réf.1.) expérience menée suivant le plan ci-après, qui se prête bien à l'analyse de variance:

- (1) La zone urbaine de Téhéran a été divisée en 21 strates, dont chacune contenait à peu près 200 pâtés de maisons ou ilôts.
- (2) Dans chaque strate, 10 ilôts ont été désignés par tirage au sort.
- (3) Parmi les ménages vivants dans chaque ilôt ainsi désigné, nous avons enquêté auprès de 5 ménages, prélevés par tirage au sort.

Les enseignements de cette enquête pilote, qui ont servi à la préparation du plan de sondage proprement dit, sont les suivantes:

a) l'analyse de variance appliquée aux données concernant chaque ménage fournit pour la plupart des rubriques du questionnaire: une variance entre les îlots à l'intérieur des strates, dont la valeur n'est pas significative;
une variance entre les strates dont la valeur est significative.

Mais d'autre part nous avons constaté que la composante de la variance, représentant la variation du nombre des ménages d'un îlot à l'autre, était si grande que, pour aboutir à un sondage suffisamment précis, il était nécessaire de renoncer à limiter l'enquête à un échantillon d'îlots.

b) En Iran, l'unité de sondage qui semble la plus commode pour un sondage concernant les ménages, n'est pas comme en France le logement ou la maison, mais la porte de cette maison; car c'est une unité qui peut être utilisée par les enquêteurs sans qu'ils aient à interroger les habitants de la maison. A l'heure actuelle, les habitants n'aiment pas être dérangés par les employés de l'administration, même dans un but purement statistique. Malheureusement nous avons rencontré une difficulté assez importante lorsque nous avons voulu adopter la porte d'immeuble comme unité de sondage: c'est qu'il existe des maisons ayant 2 portes (ou même davantage) donnant sur rue. De telles maisons auront plus de chance que les autres d'être choisies.

En fait, chaque maison aura une probabilité d'être choisie proportionnelle au nombre de ses portes (donnant sur rue). On a surmonté cette difficulté par un procédé classique (dont on ne reparlera d'ailleurs pas dans la thèse).

2. Plan proposé pour la future enquête par sondage:

a) Considérons la strate $n_{\underline{0}i}$, formée d'ilôts contigus. Soit f_{ij} le nombre des ménages qui demeurent dans la maison dont la porte aurait le $n_{\underline{0}ij}$ (dans une numération fictive); soit p_{ij} le nombre de portes (donnant sur rue) de cette maison ; soit enfin x_{ijk} l'observation x recueillie sur le ménage $n_{\underline{0}ijk}$ (par exemple le nombre de personnes composant ce ménage).

Avec ces notations, le total cherché, par exemple le nombre d'habitants de la strate $n_{\underline{0}i}$, sera

$$X_i = \sum_j \frac{f_{ij} \sum_{k=1}^{f_{ij}} x_{ijk}}{p_{ij}}$$

(où $\sum$ désigne la sommation étendue à tous les éléments de la population ; tandis que nous désignerons à l'aide du symbole S la sommation limitée aux éléments de l'échantillon). Posons:

$$y_{ij} = \frac{\sum_{k=1}^{f_{ij}} x_{ijk}}{p_{ij}}$$

et

$$Y_i = \sum_j y_{ij}$$

(1) On a bien $Y_i = X_i$

b) On se propose de soumettre à l'enquête une porte sur r (r étant un entier). Si chaque strate comprenait un nombre de portes multiple entier de r , il n'y aurait pas de problème; mais il n'y a évidemment aucune raison pour qu'il en soit ainsi. Soit M_i le nombre de portes (donnant sur la rue) que compte la strate n_i ; divisons M_i par r

$$M_i = r [m_i] + a_i$$

$$(2) \quad 0 \leq a_i < r \quad [m_i] \text{ quotient exact et } a_i \text{ reste de la division.}$$

Il est clair qu'il faut ~~prélever~~ dans la strate en échantillon de portes dont l'effectif μ_i est $[m_i]$ ou $[m_i] + 1$ suivant les strates pour avoir, dans l'ensemble, des estimateurs qui ne soient systématiquement ni trop petits, ni trop grands quand on multipliera par r les données de l'échantillon pour estimer les totaux relatifs à la population.

Voici ce que nous avons imaginé pour cela : soit V_i un nombre entier tiré au sort parmi les entiers 1 à r (inclus), tous également possibles. Nous conviendrons de prendre :

$$\mu_i = [m_i] + 1, \text{ si } V_i \leq a_i, \text{ et}$$

$$\mu_i = [m_i], \text{ si } V_i > a_i$$

Il est facile de voir que, si l'on pose $\xi_i = a_i/r$: (3)
la probabilité d'avoir $\mu_i = [m_i] + 1$ est égale à ξ_{ij}
et que la probabilité d'avoir $\mu_i = [m_i]$ est égale à $1 - \xi_i$.

Dans la strate n_{2i} , on va donc tirer au sort μ_i portes
donnant sur la rue.

c) Tous les ménages demeurant dans les maisons dont, au moins,
une porte a été ainsi désignée, seront alors soumis à l'enquête;
et les quantités y_{ij} correspondant à chacune de ces portes se-
ront finalement calculées.

Si par hasard il s'est trouvé 2 portes (ou davantage) de la
même maison parmi les μ_i portes désignées par le sort, on
affectera à chacune de ces deux portes la même quantité y_{ij} .

3. Estimation du Total.-

$$\text{Soit } Y = \sum_i y_i$$

le total cherché pour la population, par exemple le nombre total
des habitants de Téhéran. Une estimation sans biais* de Y sera

$$\begin{aligned}\hat{y} &= r \sum_j s_j y_{ij} \\ &= r \sum_j \mu_j \frac{s_j y_{ij}}{\mu_j}\end{aligned}$$

Estimation absolument convergente en probabilité vers la
quantité à estimer (traduction de l'anglais: "unbiased").

où S désigne une sommation étendue aux Y_{ij} de l'échantillon des portes.

En effet, si nous désignons l'espérance mathématique de Y par $E(\hat{Y})$, nous posons

$$n_i = \frac{y_i}{M_i}$$

et nous avons d'abord:

$$E(\hat{Y}) = r \sum_i M_i \frac{y_i}{M_i} = r \sum_i M_i n_i$$

Puis

$$\begin{aligned} E E(\hat{Y}) &= r \sum_i n_i E(\mu_i) \\ &= r \sum_i n_i [z_i ([m_i] + 1) + (1 - z_i) [m_i]] \end{aligned}$$

Appliquant les formules (2) et (3) il vient:

$$\begin{aligned} E(\hat{Y}) &= r \sum_i n_i \left[\frac{z_i}{r} \left(\frac{M_i - z_i}{r} + 1 \right) + \left(1 - \frac{z_i}{r} \right) \frac{M_i - \bar{a}_i}{r} \right] \\ &= r \sum_i n_i \frac{M_i}{r} = \sum_i M_i n_i \\ &= \sum_i y_i = Y \end{aligned}$$

4. Estimation de la Variance de $\hat{Y}$. $V(\hat{Y})$

a) Considérons un système fixe de valeurs μ_i , et calculons l'espérance mathématique $\mathcal{E}$ et la variance V de l'estimation $\hat{Y}$

$$\hat{y} = r \sum_i \mu_i \left[\frac{s_{ij} y_{ij}}{\mu_i} \right]$$

est l'estimation classique d'un sondage stratifié; de sorte que

$$\mathcal{E}(\hat{y}) = r \sum_i \mu_i \mathcal{E} \left[\frac{s_{ij} y_{ij}}{\mu_i} \right] = r \sum_i \mu_i \frac{y_i}{M_i} = y^*$$

$$\mathcal{E}(\hat{y} - y^*)^2 = v(\hat{y}) = r^2 \sum_i (\mu_i)^2 v \left[\frac{s_{ij} y_{ij}}{\mu_i} \right]$$

$$= r^2 \sum_i \mu_i^2 \frac{\sigma_i^2}{\mu_i} \frac{M_i - \mu_i}{M_i - 1}$$

$$= r^2 \sum_i \frac{\sigma_i^2}{M_i - 1} \mu_i (M_i - \mu_i)$$

b) Considérons alors l'ensemble de toutes les valeurs possibles des μ_i aléatoires et prenons l'espérance mathématique $\mathcal{E}(y^*)$ de Y^* devenu aléatoire

$$b_1) E(\hat{y}) = \mathcal{E}'[\mathcal{E}(\hat{y})] = r \sum_i \frac{y_i}{M_i} E(\mu_i) = r \sum_i \frac{y_i}{M_i} \frac{M_i}{r} = \sum_i y_i = y$$

ce qui correspond au calcul précédent (3).

$$\begin{aligned}
 b_2) \quad V(\hat{y}) &= E(\hat{y} - y)^2 = E(\hat{y} - y^* + y^* - y)^2 \\
 &= E(\hat{y} - y^*)^2 + E(y^* - y)^2 + 2E(y^* - y)(\hat{y} - y^*) \\
 &= E[V(\hat{y})] + E(y^* - y)^2 + 2E(y^* - y)[E(\hat{y} - y^*)]
 \end{aligned}$$

On voit immédiatement que le double produit est nul, et il vient:

$$(4) \quad V(\hat{y}) = r^2 \sum_i \frac{\sigma_i^2}{M_i - 1} \left[M_i E(P_i) - E(P_i^2) \right] + r^2 \sum_i \left(\frac{y_i}{M_i} \right)^2 \left[E(P_i^2) - \left(\frac{M_i}{r} \right)^2 \right]$$

Pour simplifier la formule (4) rappelons-nous que (voir formule 3)

$$P_i = \begin{cases} [m_i] + 1 = M_i/r + 1 - \varepsilon_i \\ [m_i] = M_i/r - \varepsilon_i \end{cases} \quad \text{avec la probabilité}$$

Reprenons

$$\begin{aligned}
 E(P_i) &= \varepsilon_i \left[\frac{M_i}{r} + (1 - \varepsilon_i) \right] + (1 - \varepsilon_i) \left[\frac{M_i}{r} - \varepsilon_i \right] \\
 &= \frac{M_i}{r}
 \end{aligned}$$

Donc
$$E(\mu_i) = E\left(\frac{M_i}{r}\right) = \frac{M_i}{r} \quad (5)$$

$$E(\mu_i^2) = E\left[\left(\frac{M_i}{r}\right)^2 + \varepsilon_i(1-\varepsilon_i)\right] = \left(\frac{M_i^2}{r^2}\right) + E(\varepsilon_i) - E(\varepsilon_i^2) \quad (6)$$

ε_i est une variable aléatoire qui a une distribution rectangulaire avec les limites zéro et un, indépendante de M_i et Y_i .

Donc

$$E(\varepsilon_i) = \int_0^1 t dt = \frac{1}{2}$$

$$E(\varepsilon_i^2) = \int_0^1 t^2 dt = \frac{1}{3} \quad (7)$$

Substituons (7) dans (6) et puis (5) et (6) dans (4), nous aurons en résumé:

$$\begin{aligned} V(\hat{Y}) &= r \sum_i \frac{M_i^2}{M_i-1} \frac{\sigma_i^2}{r} - r \sum_i \frac{M_i^2}{M_i-1} \frac{\sigma_i^2}{r^2} - r^2 \sum_i \frac{\sigma_i^2}{M_i-1} \frac{1}{6} + r^2 \sum_i \left(\frac{Y_i}{r}\right)^2 \frac{1}{6} \\ &= (r-1) \sum_i \frac{M_i^2 \sigma_i^2}{M_i-1} + \frac{r^2}{6} \sum_i \left[\left(\frac{Y_i}{M_i}\right)^2 - \frac{\sigma_i^2}{M_i-1} \right] \end{aligned}$$

On peut faire apparaître $\sum_i \sum_j Y_{ij}^2$ au lieu de $\sum_i Y_i^2$ ce qui présente l'avantage d'introduire une expression qui ne dépend pas du choix des strates (comme en dépendent σ_i^2 et Y_i).

On a:

$$\frac{Y_i^2}{M_i^2} = \sum_j \frac{Y_{ij}^2}{M_i} - \sigma_i^2$$

Et il vient:

$$V(\hat{y}) = (r-1) \sum_i \frac{M_i^2 \sigma_i^2}{M_i - 1} + \frac{r^2}{6} \left[\sum_i \sum_j \frac{y_{ij}^2}{M_i} - \sum_i \sigma_i^2 \frac{M_i}{M_i - 1} \right] \quad (8)$$

En supposant enfin que:

M est le nombre total des portes de Téhéran

N " " " " d'ilôts de Téhéran

L " " " " de strates de Téhéran

q " " " " d'ilôts par strate (suppose le même

pour toutes les strates)

$\sigma_i^2 = \sigma^2$ est le même dans toutes les strates

$$M_i = \frac{M_i}{L} ; q = \frac{N}{L} ; \mu_2 = \sum_i \sum_j \frac{y_{ij}^2}{M_i}$$

il vient en première approximation

$$V(\hat{y}) = (r-1) M \sigma^2 + \frac{r^2}{6} \left(\frac{\mu_2 - \sigma^2}{q} \right) \quad (9).$$

5. Le Problème de Nombre Optimum de Strates.-

1) Le calcul qui précède ne suppose rien sur la nature et le nombre de strates, jusqu'à la formule (8) incluse. On a supposé ensuite que les strates étaient de tailles égales aussi, quant au nombre des ilôts que chacune renferme.

Le problème qui se pose alors en pratique est celui du choix du plan optimum de sondage, qu'on traduit habituellement comme

Trouver le plan de sondage qui, pour un coût donné, donne la plus petite variance.

Ici le coût de l'enquête est manifestement fonction de r : un sondage au centième ($r=100$) coûtera moins cher qu'un sondage au vingtième ($r=20$). Et il est naturel de supposer ce coût indépendant du nombre de strates L , donc du nombre q d'ilôts par strate. D'où l'énoncé suivant du problème:

Pour une valeur donnée de r , trouver la valeur de q rendant $V(\hat{Y})$ minimum.

2) Toutefois une remarque de bon sens s'impose:

Si chaque îlot avait un nombre de portes M_i multiple entier de r , il est clair intuitivement qu'on aurait intérêt à considérer chaque îlot comme une strate, c'est-à-dire à élever au maximum le nombre L de strates. C'est parce que r ne divise pas exactement tous les M_i , qu'on n'a plus pour variance l'expression classique

$$\sum_i r \frac{M_i}{r} \sigma_i^2 \frac{M_i - M_i/r}{M_i - 1} = (r-1) \sum_i \frac{M_i^2 \sigma_i^2}{M_i - 1} + (r-1) M \sigma^2$$

mais une expression qui comprend

un terme principal, identique au précédent,

un terme complémentaire.

En pratique μ_2 est supérieur à σ^2 et le terme complémentaire est positif: l'irrégularité des îlots accroît la variance. Quand q diminue, c'est-à-dire quand le nombre de strates augmente, σ^2 diminue et avec lui le terme principal; mais $1/q$

et $(\mu_2 - \sigma^2)$ augmentent, de sorte que le terme complémentaire augmente. On conçoit donc l'existence d'un optimum pour une valeur de q supérieure à 1 pour laquelle $V(\hat{Y})$ sera minimum. Avec des strates trop petites l'irrégularité des îlots joue un rôle primordial et fait croître la variance $V(\hat{Y})$.

L'optimum correspond donc à une valeur de q , racine de l'équation

$$\frac{dV}{dq} = (r-1)M \frac{d\sigma^2}{dq} - \frac{r^2 N}{6q^2} - \frac{r^2 N(\mu_2 - \sigma^2)}{6q^2} = 0$$

6. Application Numérique.--

Les recherches de P.C. Mahalanobis (Ref. 2) et H. Fairfield Smith (Ref. 3) ont montré qu'une expression approchée assez satisfaisante de σ^2 pouvait souvent être employée, expression de la forme

$$\sigma^2 = Aq^g \quad (10)$$

où A est une constante positive et $0 < g < 1$, les paramètres A et g doivent être ajustés sur chaque population étudiée. Ceci suppose évidemment que σ^2 ne dépend que du nombre q d'îlots formant la strate.

Il vient alors:
$$\frac{d\sigma^2}{dq} = \frac{g}{q} \sigma^2$$

D'où une équation de deuxième degré en r , de la forme

$$r^2 - B(r-1) = 0,$$

donnant pour chaque valeur de q deux valeurs r ; on ne retient que la plus grande, voisine de B (car r et $r-1$ sont peu différents).

avec.

$$B = \frac{6AMg}{N} \frac{q^{g+1}}{12 - (1-g)Ag}$$

expression qui croît avec q , de même que B/q .

Comme exemple, nous donnons le résultat fourni par le calcul des données de l'enquête pilote de Téhéran, pour obtenir le nombre total des habitants. Nous avons calculé A et g par la méthode classique ; quant aux données relatives au calcul de M , nous n'avions qu'à prendre celles des questionnaires; et N se trouvait dans la référence 1.

Ce résultat est :

r	10	20	30	50	70	100
valeur optimum de q	2	3	5	6	7	10
coefficient de variation %	0,57	0,85	1,06	1,40	1,69	2,05

où les données numériques utilisées sont:

$$N = 4351 \quad , \quad M = 170,000 \quad , \quad \mu_2 = 135 \quad ,$$

$$g = 0,076 \quad \text{et} \quad A = 38,2$$

$T =$ le nombre des habitants de la ville de Téhéran $= 1.400.000.$

Nous avons donné ce plan comme introduction de cette thèse. Nous allons reprendre, au cours des quatre parties suivantes, le problème des erreurs introduites en arrondissant à un entier les valeurs théoriques obtenues pour tel ou tel effectif de sondage.

Dans la première partie nous exposons le problème général de l'arrondissement.

Dans la deuxième partie, nous décrivons pour divers plans de sondage les nouvelles formules d'estimateurs et de variances, en tenant compte des erreurs d'arrondissement.

Dans les troisième et quatrième parties nous résolvons différentes sortes de problèmes de recherche d'un optimum.

PREMIERE PARTIE.- Le problème général de l'arrondissement à un entier des valeurs théoriques obtenues pour tel ou tel effectif de sondage.

1.- Dans le cas d'une enquête par sondage, on dit que le statisticien est maître de son plan s'il peut contrôler toutes les sources d'erreurs. En fait, on peut prendre comme indice de la mesure du développement de la théorie de sondage, à chaque époque, le nombre de sources d'erreurs qu'on connaît et qu'on peut contrôler. Car chaque erreur qui est bien étudiée et connue - peut être évitée; ou du moins, on pourra rendre minimum son effet. Une des sources d'erreur dont l'étude est le sujet de cette thèse est l'erreur d'arrondissement des effectifs théoriques de sondage.

1°) Prenons, par exemple, le cas le plus simple : on se propose de tirer au sort un échantillon d'un ménage sur r , dans une population comprenant N ménages, où N n'est pas nécessairement un multiple entier de r et peut avoir été calculé d'après quelque formule de "l'optimum". Si N était toujours un multiple entier de r , il n'y aurait pas de problème d'erreur d'arrondissement dans notre plan. Mais si nous voulons appliquer les résultats des calculs de notre plan optimum, il n'y a évidemment aucune raison pour qu'il en soit ainsi. En fait, nous avons toujours:

$$N = r[n] + a, \quad 0 \leq a < r \quad (1)$$

où $[n]$ est le quotient et a le reste de la division.

Il est clair qu'il faut prélever dans la population un échantillon de ménages dont l'effectif, que nous désignons par μ est $[n]$ ou $[n] + 1$. (1)

Les problèmes sont ici de trouver une méthode pour choisir entre $[n]$ et $[n] + 1$, de façon que dans l'ensemble des cas possibles, la contribution des erreurs d'arrondissement à la variance des estimations soit minimum, (2) de calculer la contribution mentionnée, et (3) de modifier les formules de "l'optimum" dans lesquelles on n'a pas tenu compte des erreurs d'arrondissement.

2°) Un autre exemple (un peu plus compliqué) est le cas d'une stratification des ménages en N strates, où la strate $n^o i$ contient N_i ménages; et d'après quelque "plan optimum", on se propose de soumettre à l'enquête, dans chaque strate, un ménage sur r . Ici nous avons pour la strate $n^o i$ la formule suivante qui est l'analogue de la formule (1).

$$N_i = r [n_i] + a_i, \quad 0 \leq a_i < r$$

Il faut prélever dans la strate $n^o i$ un échantillon de ménages dont l'effectif μ_i soit $[n_i]$ ou $[n_i] + 1$ suivant les strates; et alors se pose un nouveau problème outre les 3 problèmes que nous avons énoncés à propos de l'exemple précédent; car (4) nous pouvons nous intéresser à la recherche d'une méthode pour choisir entre $[n_i]$ et $[n_i] + 1$ de telle sorte que, dans l'ensemble, nos estimateurs soient sans biais (c'est-à-dire sans erreur systématique) quand on multiplie par r les données de

l'échantillon pour estimer les totaux relatifs à la population, et (5) modifier la structure de la population pour l'amener à une structure optimum, en tenant compte des erreurs d'arrondissement.

3°) Un autre exemple qui n'est pas plus compliqué que le deuxième mais qui est d'une nature différente, est le cas d'une stratification où on veut utiliser la formule optimum de Neyman, c'est-à-dire choisir n_j , (effectif de la strate n° i) proportionnel à l'écart-type correspondant.

En utilisant la formule de Neyman, on calcule comme effectif théorique de la strate n° i un nombre n_i qui n'est pas nécessairement entier.

Nous devons donc choisir entre $[n_i]$ et $[n_i] + 1$; et nos problèmes ici sont: (1) trouver une méthode pour choisir entre $[n_i]$ et $[n_i] + 1$, de telle sorte que la différence entre la variance actuelle et la variance théorique minimum calculée par la formule de Neyman soit minimum, (2) calculer cette différence et (3) modifier la formule de calcul de n_i de telle façon qu'en tenant compte des erreurs d'arrondissement la variance soit minimum.

4°) Dans des cas de sondages plus compliqués, c'est-à-dire sondages à plusieurs degrés, avec combinaison de une ou plusieurs stratifications et d'un ou plusieurs degrés de sondage dont nous allons étudier quelques exemples au cours de cette thèse, nous rencontrons les mêmes problèmes que pour les exemples ci-dessus.

Nous allons démontrer que les propriétés essentielles restent les mêmes pour tous les cas.

2.-Etude du Problème de base.

D'abord nous étudierons le cas d'une population formée d'une seule strate contenant N éléments, dans laquelle nous nous proposons de soumettre à l'enquête un élément sur r . Puis, nous étendrons les résultats obtenus à des cas plus compliqués. Nous avons donc pour cette strate, les paramètres r , $[n]$, a et : $n = N/r$ dont les variations doivent être étudiées.

Posons

$$\frac{a}{r} = \xi \quad (3)$$

Après avoir calculé $n = N/r$, on doit choisir (d'après quelque méthode convenable) $[n]$ ou $[n] + 1$ comme effectif de sondage de notre strate unique. C'est-à-dire que nous devons choisir

$$n = [n] + Y \quad \text{où } Y \text{ est égal à } 0 \text{ ou } 1 \quad (4)$$

Si par exemple on se contentait d'arrondir à l'entier le plus proche (ou toute autre façon de faire courante) on aurait ensuite des estimateurs qui ne seraient pas exactement ceux qu'on voudrait, mais il n'y aurait pas de problème.

Nous avons un problème nouveau quand nous décidons d'employer un procédé aléatoire pour choisir entre 0 et 1 pour Y .

Le plus simple est évidemment celui qui consiste à tirer à pile ou face entre $Y = 0$ et $Y = 1$ (mais nous en imaginerons ensuite quelques-autres).

Nota: Au cours de cette thèse, nous aurons plusieurs fois l'occasion de calculer 3 sortes d'espérance mathématique : (1) l'espérance mathématique pour les variations de μ , égal soit à $[n]$, soit à $[n] + 1$, nous la désignerons par $\mathcal{E}$, (2) l'espérance mathématique pour les autres variations, que nous désignerons par $\mathcal{E}'$, et (3) l'espérance mathématique pour l'ensemble de ces deux sortes de variations qui sont orthogonales l'une à l'autre; nous la désignerons par E $E = \mathcal{E}\mathcal{E}' = \mathcal{E}'\mathcal{E}$

En tenant compte des formules (3) et (4), le problème général se pose comme suit:

D'après nos calculs, nous arrivons à un nombre $[n] + \mathcal{E}$ comme effectif du sondage; tandis que, à cause des exigences pratiques, nous sommes obligé de choisir l'un des deux nombres $[n]$ ou $[n] + 1$. Nous devons donc adopter une méthode pour choisir Y égal à 0 ou 1, en d'autres termes un mode de passage de $\mathcal{E}$ à Y .
Etude du procédé de détermination de Y par tirage au sort.

On va montrer que le tirage au sort doit être organisé de telle façon que:

$$\begin{cases} \text{prob. } (Y = 0) = 1 - \mathcal{E} \\ \text{prob. } (Y = 1) = \mathcal{E} \end{cases}$$

On se propose de soumettre l'enquête 1 sur r éléments d'une population dont l'effectif est N (non divisible par r).

Quand on ne tient pas compte des erreurs d'arrondissement, l'estimateur et sa variance sont

$$\begin{aligned} \hat{X}(n) &= r \sum_{i=1}^n x_i \\ V[\hat{X}(n)] &= \frac{N^2 (r-1) \sigma^2}{N-1} \end{aligned} \quad (5)$$

Arrondissons l'effectif n à μ de telle façon que

$$\mu = \begin{cases} [n] & \text{avec la probabilité égale à } p_0 \\ [n]+1 & \text{" " " " } p_1 \end{cases} \quad (6)$$

Si on veut tenir compte des erreurs d'arrondissement, l'estimateur (5) doit être remplacé par

$$\hat{X}(\mu) = \frac{\mu}{r} \sum_{i=1}^r x_i$$

Et pour que $\hat{X}(\mu)$ soit un estimateur sans biais, nous devons avoir

$$E[\hat{X}(\mu)] = X$$

où $X = \frac{1}{r} \sum_{i=1}^r x_i$

Nous avons

$$E[\hat{X}(\mu)] = \frac{\mu}{N} \sum_{i=1}^N x_i = \frac{\mu}{N} X = X^*$$

Il faut donc que

$$X^* = X$$

et comme

$$X^* = \frac{\mu}{N} E(\mu)$$

les probabilités p_0 et p_1 doivent être telles que

$$E(\mu) = \frac{N}{r}$$

En tenant compte de (6) et, comme $p_0 + p_1 = 1$, arrivons à

$$E(\mu) = p_0 [n] + p_1 ([n]+1) = [n] + p_1$$

d'où

$$\frac{N}{r} = [n] + p_1$$

et d'après la formule (1) il vient

$$\begin{cases} p_1 = \varepsilon & (\text{ou } \varepsilon \text{ est une constante}) \\ p_0 = 1 - \varepsilon \end{cases}$$

Ce sont les mêmes probabilités que nous avons admises, par raison de bon sens, au cours de l'introduction.

Afin de calculer la variance totale de $\hat{X}(\mu)$, nous calculons d'abord $V[\hat{X}(\mu)]$ c'est-à-dire la variance de $\hat{X}(\mu)$ quand nous considérons la valeur de μ fixe

$$V[\hat{X}(\mu)] = V(\sum_{i=1}^N x_i / \mu) = [\sum_{i=1}^N (N - \mu) \sigma^2] / (N - 1)$$

Alors nous considérons toutes les valeurs possibles de aléatoires. En tenant compte du § 4b2 de l'introduction nous avons

$$V[\hat{X}(\mu)] = \varepsilon V[\hat{X}(\mu)] + V(X^*) = \varepsilon [NE(\mu) - E(\mu^2)] \sigma^2 / (N - 1) + \varepsilon \sum_{\mu} V(\mu)$$

$$(\text{Avec } \sum_{\mu} = \frac{X}{N})$$

D'après les méthodes de calcul classique

$$V(\mu) = E(\mu - \frac{N}{\mu})^2 = (1 - \varepsilon) [E(n) - \frac{N}{\mu}]^2 + \varepsilon (E(n) + 1 - \frac{N}{\mu})^2 = (1 - \varepsilon) \varepsilon$$

d'où

$$E(\mu^2) = V(\mu) + E^2(\mu) = (1 - \varepsilon) \varepsilon + (N / \varepsilon)^2$$

Substituons les dernières formules dans (8), nous avons

$$V[\hat{X}(\mu)] = N \frac{\varepsilon - 1}{N - 1} \sigma^2 + \frac{\varepsilon}{N - 1} \varepsilon (1 - \varepsilon) [\sum_{\mu} - \sigma^2 / (n - 1)]$$

On voit qu'ici aussi, comme dans le cas étudié dans l'introduction, $V[\hat{X}(M)]$ comprend deux termes dont le premier est $V[\hat{X}(n)]$ et le deuxième est la contribution des erreurs d'arrondissement; et il est d'autant plus grand que $|\varepsilon - \frac{1}{2}|$ est plus petite.

3- Cas de plusieurs strates.

Prenons d'abord le cas où nous avons deux strates. Dans ce cas, (et aussi dans les cas où nous avons plus de 2 strates) on peut adopter deux façons de choisir égal à $[n]$ ou $[n]+1$: choix indépendants ou choix non indépendants. Dans le cas des choix indépendants on arrondit les n_i dans chaque strate indépendamment des choix faits dans les autres strates, ce qui n'entraîne pas de problème nouveau. Les formules correspondantes sont:

$$\left[\begin{aligned} \hat{X}(n) &= \varepsilon \sum_{k=1}^N S_k x_{k\varepsilon} \\ V[\hat{X}(n)] &= (\varepsilon - 1) \sum_{k=1}^N \frac{N_k^2 \varepsilon_k^2}{N_k - 1} \quad (N=2) \quad (10) \\ V[\hat{X}(M)] &= V[\hat{X}(n)] + \varepsilon \sum_{k=1}^N \left[\varepsilon_k (1 - \varepsilon_k) \varepsilon_k^2 - \frac{\varepsilon_k^2}{N_k - 1} \right] \end{aligned} \right.$$

Pour le cas des choix non indépendants de μ_1 , on va étudier à titre d'exemple le cas où N est exactement divisible par r . La meilleure méthode semble être de choisir $\mu_2 = n - \mu_1$ après avoir arrondi l'effectif de la strate 1 à μ_2 . Nous avons $E(\mu_2) = E(n - \mu_1) = n - E(\mu_1) = n - n_1 = n_2$

Il est évident que

$$V(\mu_2) = V(\mu_1) = \sum_i (1 - \varepsilon_i) \quad (11)$$

Tout d'abord

$$V[\hat{X}(\mu)] = V(\sum_i \varepsilon_i x_{1i}) + V(\sum_i \varepsilon_i x_{2i}) + 2 \text{Cov}(\sum_i \varepsilon_i x_{1i}, \sum_i \varepsilon_i x_{2i})$$

La covariance est nulle, car nous examinons le cas de deux tirages indépendants (avec des effectifs μ_1 μ_2 fixes) dans deux strates.

Il reste donc

$$V[\hat{X}(\mu)] = \frac{\sum_i \mu_i (N_i - \mu_i) \sigma_i^2}{N_1 - 1} + \frac{\sum_i \mu_i (N_i - \mu_i) \sigma_i^2}{N_2 - 1}$$

Faisons à présent varier μ_1 et μ_2 . Il vient

$$V[\hat{X}(\mu)] = \sum V[\hat{X}(\mu)] + V(X^*) \quad (12)$$

avec

$$X^* = \sum (\frac{\mu_1}{N_1} X_1 + \frac{\mu_2}{N_2} X_2)$$

$$V(X^*) = V(X_1^*) + V(X_2^*) + 2 \text{Cov}(X_1^*, X_2^*)$$

$$\text{Cov}(X_1^*, X_2^*) = \sum (X_1^* - X_1)(X_2^* - X_2)$$

$$= \sum (\frac{\sum \mu_i X_i}{N_1} - X_1)(\frac{\sum \mu_i X_i}{N_2} - X_2)$$

$$= \sum (\frac{\sum \mu_i X_i}{N_1} - \frac{\sum n_i X_i}{N_1})(\frac{\sum \mu_i X_i}{N_2} - \frac{\sum n_i X_i}{N_2})$$

$$= \sum \frac{\sum X_i X_j}{N_1 N_2} (\mu_i - n_i)(\mu_j - n_j)$$

$$= \sum_i \sum_j \varepsilon_{ij} \text{Cov}(\mu_i, \mu_j)$$

ce qui donne, en tenant compte de la formule (11) et du fait que la corrélation entre μ_1 et μ_2 est - 1,

$$\begin{aligned} \text{Cov}(X_1^*, X_2^*) &= \varepsilon^2 \xi_1 \xi_2 \sqrt{V(\mu_1) V(\mu_2)} (-1) \\ &= -\varepsilon^2 \xi_1 \xi_2 \varepsilon_1 (1 - \varepsilon_1) \end{aligned}$$

Et comme

$$V(X_K^*) = \varepsilon^2 \xi_{K0}^2 V(\mu_K) = \varepsilon^2 \xi_{K0}^2 \varepsilon_1 (1 - \varepsilon_1)$$

(K = 1 et 2)

nous avons:

$$V(X^*) = \varepsilon^2 \left(\xi_{10}^2 + \xi_{20}^2 - 2 \xi_{10} \xi_{20} \right) \varepsilon_1 (1 - \varepsilon_1) = \varepsilon^2 \varepsilon_1 (1 - \varepsilon_1) (\xi_{10} - \xi_{20})^2 \quad (13)$$

D'autre part, en faisant les mêmes sortes de calculs que pour la formule (9), et en tenant compte des formules (10) et (11), nous avons

$$E'V[\hat{X}(\mu)] = V[\hat{X}(n)] - \varepsilon^2 \sum_{K=1}^2 \frac{\sigma_K^2}{N_K^{-1}}$$

Substituons (13) et la dernière formule dans (12); nous obtiendrons

$$V[\hat{X}(\mu)] = V[\hat{X}(n)] + \varepsilon^2 \varepsilon_1 (1 - \varepsilon_1) \left[(\xi_{10} - \xi_{20})^2 - \sum_{K=1}^2 \frac{\sigma_K^2}{N_K^{-1}} \right]$$

Afin de comparer les résultats des deux sortes de choix possibles pour μ_i , posons :

V (choix indépendants) - V (choix non indépendants) = D , et supprimons les termes contenant $\sigma_{k/}^2 (N_k - 1)$ qui sont négligeables; nous avons

$$D = r^2 \left\{ \varepsilon_1 (1 - \varepsilon_1) \left[\bar{\xi}_{10}^2 - (\bar{\xi}_{10} - \bar{\xi}_{20})^2 \right] + \varepsilon_2 (1 - \varepsilon_2) \bar{\xi}_{20}^2 \right\}$$

$$= r^2 \left[2 \bar{\xi}_{10} \bar{\xi}_{20} \varepsilon_1 (1 - \varepsilon_1) + \bar{\xi}_{20}^2 (\varepsilon_1 - \varepsilon_1) (1 - \varepsilon_1 - \varepsilon_2) \right]$$

On voit que, si $\varepsilon_1 = \varepsilon_2$, on a toujours $D > 0$, et si

$\varepsilon_1 \neq \varepsilon_2$, on peut toujours s'arranger par choisir l'une ou l'autre des strates comme strate 1, de telle façon que D soit positive; ainsi le choix non indépendant est plus efficace que l'autre.

Quand nous avons plus de 2 strates, le cas du choix indépendant des μ_i est encore simple. Les formules correspondantes sont les formules (10) avec $N > 2$. Mais si on veut procéder à des choix non indépendants, la méthode que nous avons utilisée devient très compliquée. Nous procédons donc à la méthode de report du reste à l'unité suivante, que nous allons expliquer ci-dessous.

On construit la table des fréquences cumulées des strates, à condition que les strates soient rangées par tirage au sort dans la table. Sur la ligne n°i de la table on écrit la fréquence

$$F_i = \sum_{t=1}^i N_t$$

Puis on tire au sort un nombre entier entre 1 et r (inclus), par exemple s, avec les probabilités égales. On fait correspondre le nombre entier $s + kr$ ($k =$ entier quelconque) à la strate $n^o i$ à condition que $F_{i-1} \leq s + kr \leq F_i$. Dans chaque strate le nombre des éléments qui doivent être soumis à l'enquête, c'est-à-dire N_i sera égal au nombre des entiers $s + kr$ (pour toute valeur k satisfaisant à la double inégalité ci-dessus) qui correspondent à cette strate.

Il nous semble que, dans le cas où nous avons plusieurs strates, la méthode du reporté du reste est la meilleure; car il est intuitivement évident que les restes des différentes strates se compensent d'une façon beaucoup plus homogène que, par exemple, la méthode d'arrondissement à l'entier le plus proche. Mais le traitement mathématique de cette méthode, pour le cas où le nombre des strates est > 2 , mais petit, est très compliqué; et nous allons nous contenter de reprendre cette méthode avec les autres pour le cas où les strates sont nombreuses.

4-Cas d'un plan impliquant de nombreux arrondissements.

Lorsqu'on doit procéder à de nombreux arrondissements, il n'est plus du tout utile de tirer vraiment au sort les valeurs Y . On peut admettre en effet que les ξ correspondants (s'ils sont nombreux) sont les valeurs prises par une véritable variable aléatoire.

On désignera dans la suite la variable aléatoire dont la loi de distribution est rectangulaire entre les limites (a,b)

par $\text{rect } (a,b)$.

Ainsi, on a pratiquement intérêt à admettre que

$$\xi = \text{rect } (0,1)$$

Ouvrons une parenthèse. Considérons, par exemple, comme dans l'introduction, une population formée de nombreuses strates d'effectifs N_i . L'entier r est un paramètre et on considère en revanche que, pour une strate donnée, N_i est aléatoire et

$$\text{rect } [r [n_i], r [n_i] + r]$$

donc que n_i est aléatoire et

$$\text{rect } [[n_i], [n_i] + 1]$$

Cela peut paraître surprenant à première vue, mais est au fond très naturel. C'est ainsi que quand on dit que la ville de Paris a (en 1954) 2.850.189 habitants, il est bien clair que les dernière décimales sont plus ou moins aléatoires (Rappelons que Tippett a fabriqué une table de nombres aléatoires en recopiant de telles décimales sur le volume de Recensement anglais).

En tout cas il est bien évident que, si la population N_i venait à être du même ordre de grandeur que r , la variable n_i pourrait présenter des caractères pathologiques (être plus petite que 1). On va, dans la suite, supposer que n_i n'est en tous cas pas inférieure à 1 (c'est-à-dire ne peut pas être arrondie à 0).

On va à présent énumérer un certain nombre de procédés aléatoires que nous emploierons, pour "arrondir" nos effectifs. Nous n'étudierons aucun procédé "biaisé", c'est-à-dire créant une erreur systématique comme, par exemple, celui qui consiste à arrondir toujours à l'entier inférieur.

Les méthodes "sans biais" que nous étudierons sont:

La méthode A-arrondissement à l'entier le plus proche

La méthode A'- " " " éloigné

La méthode A''- " " à un des entiers $[n]$ ou $[n]+1$ par tirage au sort avec probabilités égales.

La méthode B-arrondissement à l'un des entiers $[n]$ et $[n]+1$ par tirage au sort, avec respectivement les probabilités $1-\varepsilon$ et ε .

La méthode C-la méthode de report du reste à l'unité suivante.

Nous allons d'abord expliquer ces méthodes dans le cas où les strates sont très nombreuses.

Méthode A- Après avoir calculé $n = N/r$, on arrondit n à son entier le plus proche. Dans ce cas $\mu = \begin{cases} [n]+1 & \text{avec la probabilité } 1/2 \\ [n] & \text{avec la probabilité } 1/2 \end{cases}$

Méthode A'- Après avoir calculé $n = N/r$, on arrondit n à l'entier le plus éloigné. Dans ce cas μ est encore donné par la formule précédente.

Méthode A''- Après avoir calculé $n = N/r$, on choisit $[n]$ ou $[n]+1$ par tirage au sort, avec les probabilités égales. Ici la dernière formule donne encore μ .

La méthode B- Cette méthode est la même que celle utilisée au cours de l'introduction (voir § 2 de l'introduction).

La méthode C- Nous avons introduit déjà cette méthode au § 3. Donc nous donnons ici des explications complémentaires.

Supposons que le nombre $s+k_1r$ soit le dernier qui tombe dans la fréquence cumulée n° $i-1$, et par conséquent que le nombre $s+(k_1+1)r$ soit le premier qui tombe dans la fréquence cumulée n° i . Supposons qu'après être arrivé à -

l'élément de la strate n° $i-1$ qui correspond au nombre $s+k$ if , il est resté $b_{i-1} = F_{i-1} - (s+k-1-r)$ ($0 \leq b_{i-1} < r$) = nombre des éléments de la strate n° $i-1$ qui doivent être reportés à la strate n° i pour être complétés par $r-b_{i-1}$ des éléments de la strate i , et (après avoir été additionné à $s+k-1-r$) constituer le 1° nombre $s+(k+1)$ qui corresponde à la strate n° i . Posons $r-b_{i-1} = \varepsilon_i$. Il est clair que η_i est une variable aléatoire. Nous avons donc (avec cette méthode) deux variables, η et ε (de même que dans le cas de la méthode B nous avons deux variables ν et ε). Nous étudierons la distribution de η_i au cours du § 6.

Si, dans la strate n° i , on a $\eta_i \leq \varepsilon_i$, le nombre des éléments qui seront tirés de cette strate sera, manifestement, $[\eta_i] + 1$ et il reste encore de cette strate $b_i = r(\varepsilon_i - \eta_i)$ éléments qui seront reportés sur la strate n° $i+1$ et par conséquent $\eta_{i+1} = 1 + (\eta_i - \varepsilon_i)$.

Si $\eta_i > \varepsilon_i$, le nombre des éléments qui seront soumis à l'enquête pour la strate n° i est $[\eta_i]$ et, il est clair que $\eta_{i+1} = \eta_i - \varepsilon_i$

Si nous désignons le nombre des éléments qui seront choisis dans la strate n° i par μ_i , la fonction de densité de η_i par $f_i(\eta)$ et $\int_0^x f_i(\eta) d\eta$ par $F_i(x)$, nous avons

$$\mu = \begin{cases} [\eta_i] + 1 & \text{avec la probabilité } F_i(\eta_i) \\ [\eta_i] & \text{avec la probabilité } 1 - F_i(\eta_i) \end{cases}$$

De même que ξ , on peut admettre que η est rect (0,1) et aussi que ξ et η sont indépendants l'un de l'autre et indépendants des autres paramètres.

Dans les deux § qui suivent, nous allons cependant essayer de démontrer que, sous certaines conditions assez peu restrictives, on aura bien

$$\xi = \text{rect} (0,1) \quad , \quad \eta = \text{rect} (0,1)$$

On peut aussi bien, d'ailleurs, postuler ces propriétés dans la suite du présent exposé.

5 - Distribution de ξ . Rappelons que ξ est défini par l'identité

$$N = r [n] + a, \quad 0 \leq a < r \quad \text{et} \quad \xi = \frac{a}{r}$$

Nous allons étudier la distribution de ξ pour l'une des strates. On pourra vérifier, facilement, que le résultat est le même pour l'ensemble des strates.

Pour une strate donnée, dont le nombre des éléments N serait donné avec r fixe, ξ serait aussi fixe, et nous n'aurions plus aucun problème. Nous allons donc admettre que $n = N/r$ est une variable aléatoire continue tandis que r est un paramètre continu. Nous appellerons la loi de la distribution de n , $f(n) dn$. Nous n'avons pas besoin de connaître $f(n)$. Mais nous posons comme condition qu'elle a un contact d'ordre assez grand, avec l'axe des abscisses, en ses deux extrémités. Cette condition qui a été imposée par Sheppard aux distributions statistiques, afin de démontrer ses formules de corrections pour le groupement des données, a déjà été vérifiée pour la plupart

des distributions statistiques réelles et semble être raisonnable pour le cas de $f(n) dn$.

Ecrivons la formule (1) sous la forme

$$n = [n] + \varepsilon$$

ce qui donne

$$dn = d\varepsilon$$

et par conséquent, si nous désignons la loi de distribution de

ε par $dF(\varepsilon)$, nous avons

$$d f(\varepsilon) = f([n] + \varepsilon) d\varepsilon = \sum_{v=0}^{\infty} f^{(v)}([n]) \frac{\varepsilon^v}{v!} d\varepsilon$$

Le moment d'ordre s de ε est donc

$$\mu_s(\varepsilon) = \sum_{[n]=0}^{\infty} \int_0^1 \varepsilon^s \sum_{v=0}^{\infty} f^{(v)}([n]) \frac{\varepsilon^v}{v!} d\varepsilon$$

Comme les conditions pour échanger les places de $\int$ et des $\sum$ sont remplies et

$$\int_0^1 \varepsilon^{s+v} d\varepsilon = \frac{1}{s+v+1}$$

nous avons

$$\begin{aligned} \mu_s(\varepsilon) &= \sum_{[n]=0}^{\infty} \sum_{v=0}^{\infty} f^{(v)}([n]) \frac{1}{v!} \int_0^1 \varepsilon^{v+s} d\varepsilon \\ &= \sum_{v=0}^{\infty} \frac{1}{v!(s+1)} \sum_{[n]=0}^{\infty} f^{(v)}([n]) \end{aligned}$$

Nous pouvons poser, avec une très bonne approximation $\sum_{[n]=0}^{\infty} f^{(v)}([n])$ égale à $\int_0^{\infty} f^{(v)}(n) dn$; et comme $f(n)$ est une fonction de densité avec contacts d'ordre assez grand avec l'axe des abscisses, $\int_0^{\infty} f^{(v)}(n) dn$ est égale à 0 pour $v > 0$, et égale à 1 pour $v = 0$, tandis que le facteur $1/v!$ tend rapidement vers zéro. Nous avons donc

$$\mu_s(\varepsilon) = 1/(s+1)$$

On voit que la distribution de ξ est telle que son moment d'ordre s soit $1/(s+1)$, pour tout s . D'après le théorème des moments, une distribution est complètement définie par moments des différents ordres. La distribution de ξ est donc rect(0,1).

Un corollaire immédiat de cette proposition, c'est-à-dire que avec une très bonne approximation nous pouvons admettre que ξ est indépendant des autres paramètres de la population.

Le deuxième corollaire, c'est que, dans le cas de sondages plus compliqués les ξ des différentes unités de sondage sont indépendants les uns des autres; et chacun est séparément rect (0,1)

6.- Distribution de η_i

Rappelons que nous avons défini η_i quand nous avons parlé de la méthode C, au § 4, par l'identité

$$\xi - b_{i-1} = \xi \eta_i \quad 0 \leq b_{i-1} < \xi$$

où b_{i-1} est le reste de la strate $n^{o}i-1$ qu'on reporte à la strate $n^{o} i$.

Définissons la variable aléatoire

$$\theta_{i+1} = \eta_i - \xi = \begin{cases} \eta_{i+1} - 1 & \text{si } \eta_i \leq \xi \\ \eta_{i+1} & \text{si } \eta_i > \xi \end{cases} \quad (d'où -1 \leq \theta_{i+1} \leq 1)$$

Si nous désignons la loi de distribution de θ_{i+1} par $dG_{i+1}(\theta)$

) et celle de ξ qui est rect (0,1), par $h(\xi) d\xi$,

nous avons

$$dG_{i+1}(\theta) = d\theta \int_0^1 h(\xi) f_i(\theta_{i+1} + \xi) d\xi = d\theta \int_0^1 f_i(\theta_{i+1} + \xi) d\xi$$

et d'après la définition de θ_{i+1}

$$dF_{i+1}(n) = dG_{i+1}(n) + dG_{i+1}(n-1) \quad (I4)$$

Supposons alors que $f_i(n)$ soit un polynôme de degré t , c'est-à-dire

$$f_i(n) = \sum_{v=0}^t a_v n^v$$

Nous avons, dans ce cas

$$\begin{aligned} dG_{i+1}(\theta) &= d\theta \sum_{v=0}^t a_v \int_0^1 (\theta_{i+1} + \varepsilon \varepsilon_i)^v d\varepsilon = d\theta \sum_{v=0}^t \frac{a_v [(\theta_{i+1} + 1)^v - \theta_{i+1}^v]}{v+1} \\ &= d\theta \sum_{v=0}^t b_v \theta_{i+1}^v \end{aligned}$$

où les b_v sont constantes.

D'après la formule (I4)

$$dF_{i+1}(n) = d\eta \left[\sum_{v=0}^t b_v [n^v + (n-1)^v] \right] = d\eta \sum_{v=0}^t c_v n^v$$

où c_v sont aussi des constantes.

On voit que, si $f_i(n)$ est un polynôme de degré n ,

$f_{i+1}(n)$ en est un aussi et, comme d'après la définition de la méthode C, $f_1(n)$ est pratiquement rect(0,1), c'est-à-dire un polynôme de degré 0, par récurrence, nous pouvons dire que $f_{i+1}(n)$ est aussi un polynôme de degré 0, c'est-à-dire rect(0,1). Les limites (0,1) que nous avons indiquées pour cette distribution viennent directement de la définition de η_i .

Les corrolaires que nous avons donnés à la fin de §5 peuvent être appliqués ici aussi:

Pour n'importe quelle sorte de sondage (simple, avec -

stratification, sondage à plusieurs degrés et leurs combinaisons), si nous appliquons la méthode C, les η concernant les différentes unités sont indépendants entre eux, des ξ et des autres paramètres de la population; et dans tous les cas ils sont $\text{rect}(0,1)$.

7.- Etude comparée de nos diverses méthodes d'arrondissement.

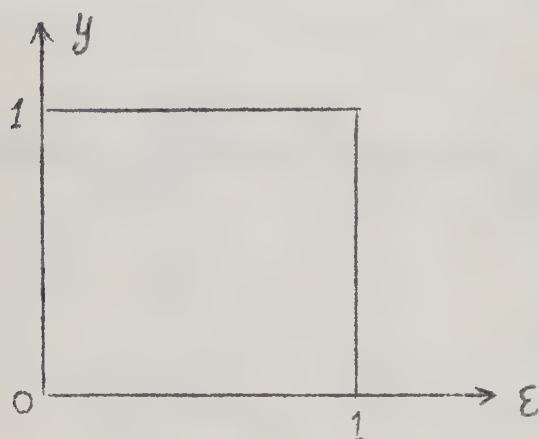


Figure 1

Puisqu'il s'agit finalement de remplacer une quantité ξ aléatoire $\text{rect}(0,1)$ par une variable aléatoire Y égale à 0 ou à 1, M. THIONET nous a suggéré de calculer pour chaque méthode l'expression

$$E(Y - \xi)^2 \text{ et } E(Y - \xi)$$

et de voir si la meilleure méthode ne correspond pas à celle qui rend la première de ces quantités nulle

et la seconde minimum.

On a donc étudié de ce point de vue les diverses méthodes; les résultats sont en accord avec l'hypothèse formulée; mais nous n'avons pas été en mesure d'établir sur ce point de résultat absolument général.

Tout d'abord il faut avoir $E(Y) = E(\xi)$, donc

$$\text{Prob}(Y = 0) = \text{prob}(Y = 1) = 1/2$$

D'autre part on a

$$\begin{aligned} E(Y - \xi)^2 &= E(Y - 1/2)^2 + E(\xi - 1/2)^2 - 2 \text{COV}(Y, \xi) \\ &= V(Y) + V(\xi) - 2 \rho_{Y\xi} \sqrt{V(Y) V(\xi)} \quad (15) \end{aligned}$$

ce qui est la même chose que la variance de μ , $V(\mu)$, car

$$E(Y - \varepsilon)^2 = E(Y + [n] - n)^2 = E(\mu - n)^2 = E[\mu - E(\mu)]^2 = V(\mu)$$

La méthode A - Manifestement $p_0 = p_1 = 1/2$. Cette méthode

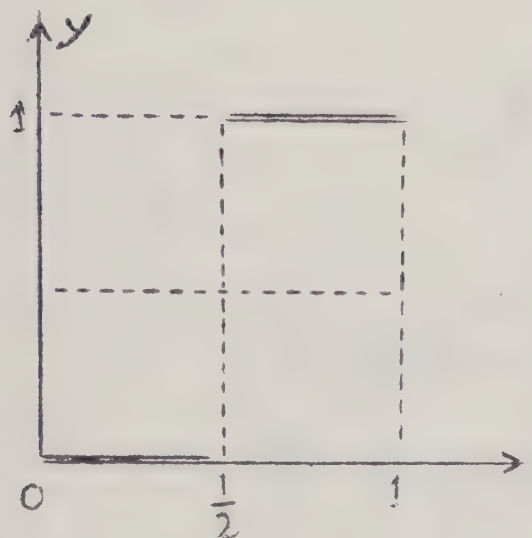


Figure 2

peut être illustrée par le schéma de la figure 2. C'est-à-dire la courbe représentant la variation de Y par rapport à ε est formée de deux parties: la première sur l'axe des abscisses entre 0 et 1/2, et la deuxième sur la droite $Y = 1$ entre les abscisses 1/2 et 1. La formule (15)

ici est donc

$$E(Y - \varepsilon)^2 = \int_0^{1/2} (0 - \varepsilon)^2 d\varepsilon + \int_{1/2}^1 (1 - \varepsilon)^2 d\varepsilon = \frac{1}{12}$$

Pour les 5 méthodes, nous avons $V(\varepsilon) = 1/3 - (1/2)^2 = 1/12$ et $V(Y) = 1/2 - (1/2)^2 = 1/4$. D'après la formule (15) nous avons

$$\frac{1}{4} + \frac{1}{12} - 2\rho_{Y\varepsilon}\sqrt{\frac{1}{4} \cdot \frac{1}{12}} = \frac{1}{12}$$

d'où

$$\rho_{Y\varepsilon} = \sqrt{3}/2$$

Pour calculer les différentes espérances mathématiques, nous considérons que pour cette méthode

$$0 \leq \varepsilon \leq \frac{1}{2}$$

$$N = z[n] + z - \frac{z}{2} + z\varepsilon$$

$$\text{quand } \mu = [n] + 1$$

$$N = z[n] + z\varepsilon$$

$$\text{" } \mu = [n]$$

Donc

$$\mathcal{E}(\varepsilon) = 2 \int_0^{\frac{1}{2}} x dx = \frac{1}{4}$$

$$\begin{aligned} E(\mu) &= \mathcal{E}[\mathcal{E}'(\mu)] = \mathcal{E}\left[\frac{1}{2}\left(\frac{N}{2} - \frac{1}{2} - \varepsilon + 1\right) + \frac{1}{2}\left(\frac{N}{2} - \varepsilon\right)\right] \\ &= \mathcal{E}\left[\frac{N}{2} - \varepsilon + \frac{1}{4}\right] \\ &= \frac{N}{2} - \frac{1}{4} + \frac{1}{4} = \frac{N}{2} \end{aligned}$$

et

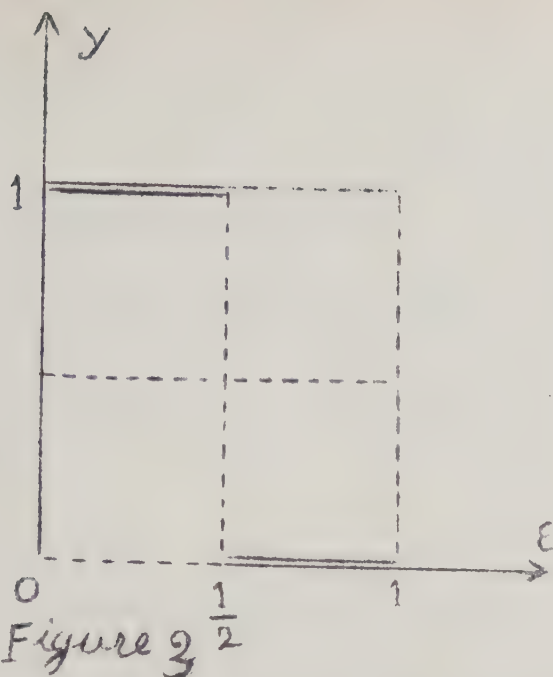
$$\mathcal{E}(\varepsilon^2) = 2 \int_0^1 x^2 dx = \frac{1}{12}$$

$$\begin{aligned} E(\mu^2) &= \mathcal{E}\left[\frac{1}{2}\left(\frac{N}{2} - \frac{1}{2}\right)^2 + \frac{1}{2}\left(\frac{N}{2} - \varepsilon\right)^2\right] \\ &= \frac{1}{2} \mathcal{E}\left[\left(\frac{N}{2}\right)^2 + \frac{1}{4} + \varepsilon^2 + \frac{N}{2} - \frac{2N\varepsilon}{2} - \varepsilon\right. \\ &\quad \left.+ \left(\frac{N}{2}\right)^2 + \varepsilon - \frac{2N\varepsilon}{2}\right] \\ &= \left(\frac{N}{2}\right)^2 + \frac{1}{8} + \frac{1}{12} + \frac{N}{2} - \frac{N}{2} - \frac{1}{8} \\ &= \left(\frac{N}{2}\right)^2 + \frac{1}{12} \end{aligned}$$

Donc

$$V(\mu) = E(\mu^2) - E^2(\mu) = \frac{1}{12}$$

La méthode A' - Ici aussi, on a manifestement $p_0 = p_1 = 1/2$. Cette méthode peut être illustrée par la figure 3 qui est la même que la figure 1 où on a interverti les deux segments.



La formule (15) ici est donc:

$$E(Y - \xi)^2 = \int_0^{\frac{1}{2}} (1 - \xi)^2 d\xi + \int_{\frac{1}{2}}^1 (0 - \xi)^2 d\xi$$

$$= \frac{7}{12}$$

donc

$$\frac{1}{4} + \frac{1}{12} - 2\rho_{y\xi}\sqrt{\frac{1}{4} \cdot \frac{1}{12}} = \frac{7}{12}$$

d'où

$$\rho_{y\xi} = -\frac{\sqrt{3}}{2}$$

Pour calculer les différentes espérances mathématiques, considérons que, dans le cas de cette méthode:

$$\frac{1}{2} \leq \xi < 1$$

$$N = r[n] + r\xi - \frac{r}{2} \text{ quand } \mu = [n] + 1$$

$$N = r[n] + r\xi \text{ quand } \mu = [n]$$

Donc

$$E(\xi) = 2 \int_{\frac{1}{2}}^1 x dx = \frac{3}{4}$$

$$E(\mu) = E[E'(\mu)] = E\left[\frac{1}{2}\left(\frac{N}{r} + \frac{3}{2} - \xi\right) + \frac{1}{2}\left(\frac{N}{r} - \xi\right)\right]$$

$$= \frac{N}{r} + \frac{3}{4} - \frac{3}{4} = \frac{N}{r}$$

et

$$E(\xi^2) = \int_{\frac{1}{2}}^1 x^2 dx = 9/12$$

$$E(\mu^2) = E[E'(\mu^2)] = E\left[\frac{1}{2}\left(\frac{N}{r} + \frac{3}{2} - \xi\right)^2 + \frac{1}{2}\left(\frac{N}{r} - \xi\right)^2\right]$$

$$= \frac{1}{2}\left[\left(\frac{N}{r}\right)^2 + \frac{9}{4} + \xi^2 + \frac{3N}{r} - \frac{3N\xi}{r} - 3\xi + \left(\frac{N}{r}\right)^2 + \xi^2 - \frac{2N\xi}{r}\right]$$

$$= \frac{1}{2}\left[2\left(\frac{N}{r}\right)^2 + \frac{9}{4} + \frac{7}{12} + \frac{3N}{r} - \frac{3N}{r} - \frac{3}{4} + \frac{7}{12} - \frac{3N}{2r}\right]$$

$$= \left(\frac{N}{r}\right)^2 + \frac{7}{12}$$

d'où

$$V(\mu) = E(\mu^2) - E^2(\mu) = \frac{7}{12}$$

La méthode A'' - Ici aussi la proposition $p_0 = p_1 = 1/2$ est évidente et $\rho_{y\varepsilon}$ est évidemment nul. La formule (15) s'écrit donc

$$E(Y - \varepsilon)^2 = \frac{1}{4} + \frac{1}{12} = \frac{1}{3}$$

Pour calculer les différentes espérances mathématiques, nous considérerons que $0 \leq \varepsilon < 1$. Nous avons

$$E(\varepsilon) = \int_0^1 x dx = \frac{1}{2}$$

$$E(\mu) = E[E'(\mu)] = E\left[\frac{1}{2}\left(\frac{N}{2} - \varepsilon + 1\right) + \frac{1}{2}\left(\frac{N}{2} - \varepsilon\right)\right] = \frac{N}{2} - \frac{1}{2} + \frac{1}{2} = \frac{N}{2}$$

$$E(\varepsilon^2) = \int_0^1 x^2 dx = 1/3$$

$$E(\mu^2) = E\left[\frac{1}{2}\left(\frac{N}{2} + 1 - \varepsilon\right)^2 + \frac{1}{2}\left(\frac{N}{2} - \varepsilon\right)^2\right]$$

$$= \frac{1}{2} E\left[\left(\frac{N}{2}\right)^2 + \varepsilon^2 + 1 + \frac{2N}{2} - \frac{2N}{2} - 2\varepsilon + \left(\frac{N}{2}\right)^2 + \varepsilon^2 - \frac{2N\varepsilon}{2}\right]$$

$$= \left(\frac{N}{2}\right)^2 + \frac{1}{3} + 1 + \frac{2N}{2} - \frac{2N}{2} - 1 = \left(\frac{N}{2}\right)^2 + \frac{1}{3}$$

Donc

$$V(\mu) = E(\mu^2) - E^2(\mu) = \frac{1}{3}$$

La méthode B - Nous pouvons démontrer que, pour cette méthode encore $p_0 = p_1 = 1/2$. En effet

$$\begin{aligned} P_1 = P(V \leq a) &= \int_0^1 d\varepsilon \int_0^\varepsilon \frac{v}{\varepsilon} d\frac{v}{\varepsilon} \\ &= \int_0^1 \varepsilon d\varepsilon = \frac{1}{2} \end{aligned}$$

Pour calculer $\rho_{y\varepsilon}$

$$\begin{aligned} \text{Cov}(y, \varepsilon) &= E\left(y - \frac{1}{2}\right)\left(\varepsilon - \frac{1}{2}\right) \\ &= \int_0^1 \left(0 - \frac{1}{2}\right)\left(\varepsilon - \frac{1}{2}\right)(1-\varepsilon) d\varepsilon + \int_0^1 \left(1 - \frac{1}{2}\right)\left(\varepsilon - \frac{1}{2}\right)\varepsilon d\varepsilon = \frac{1}{12} \end{aligned}$$

$$\rho_{y\varepsilon} = \frac{1}{12} / \sqrt{\frac{1}{4} \cdot \frac{1}{12}} = 1/\sqrt{3}$$

La formule (15) donne donc

$$E(y - \varepsilon)^2 = \frac{1}{4} + \frac{1}{12} - \frac{2}{\sqrt{3}} \sqrt{\frac{1}{4} \cdot \frac{1}{12}} = \frac{1}{6}$$

Pour calculer les différentes espérances mathématiques, considérons que dans le cas de cette méthode $0 \leq \varepsilon < 1$:

$$E(\varepsilon) = \int_0^1 x dx = \frac{1}{2}$$

$$E(\mu) = E\left[\varepsilon\left(\frac{N}{2} + 1 - \varepsilon\right) + (1 - \varepsilon)\left(\frac{N}{2} - \varepsilon\right)\right] = E\left[\frac{N\varepsilon}{2} + \varepsilon - \varepsilon^2 + \frac{N}{2} - \varepsilon - \frac{\varepsilon N}{2} + \varepsilon^2\right]$$

$$= \frac{N}{2\varepsilon} + \frac{1}{2} - \frac{1}{3} + \frac{N}{2} - \frac{1}{2} - \frac{N}{2\varepsilon} + \frac{1}{3} = \frac{N}{\varepsilon}$$

et

$$E(\varepsilon^2) = \int_0^1 \varepsilon^2 d\varepsilon = \frac{1}{3}$$

$$E(\mu^2) = E\left[\varepsilon\left(\frac{N}{2} + 1 - \varepsilon\right)^2 + (1 - \varepsilon)\left(\frac{N}{2} - \varepsilon\right)^2\right]$$

$$= E\left[\varepsilon\left(\frac{N}{2}\right)^2 + \varepsilon + \varepsilon^3 + \frac{2N\varepsilon}{2} - 2\varepsilon^2 - \frac{2N\varepsilon}{2} + \left(\frac{N}{2}\right)^2 + \varepsilon - \frac{2N\varepsilon}{2}\right.$$

$$\left. - \varepsilon\left(\frac{N}{2}\right)^2 - \varepsilon + \frac{2N\varepsilon}{2}\right] = \frac{1}{2} - \frac{1}{3} + \left(\frac{N}{2}\right)^2 = \left(\frac{N}{2}\right)^2 + \frac{1}{6}$$

Donc

$$V(\mu) = E(\mu^2) - E(\mu)^2 = \frac{1}{6}$$

Méthode C - D'abord nous démontrons que $p_0 = p_1 = 1/2$. En effet

$$T_1 = P(\varepsilon \leq \eta) = \int_0^1 d\varepsilon \int_0^\varepsilon d\eta = \frac{1}{2}$$

On voit qu'il y a une certaine ressemblance entre les méthodes B et C. On verra que cette ressemblance existe entre elles au point de vue des autres propriétés aussi. En fait, dans le cas d'une seule strate, elles sont identiques. Calculons $P_{y\varepsilon}$:

$$\begin{aligned} \cos v(y, \varepsilon) &= E\left(y - \frac{1}{2}\right)\left(\varepsilon - \frac{1}{2}\right) = \int_0^1 \left(0 - \frac{1}{2}\right)\left(\varepsilon - \frac{1}{2}\right) d\varepsilon \int_\varepsilon^1 d\eta \\ &+ \int_0^1 \left(1 - \frac{1}{2}\right)\left(\varepsilon - \frac{1}{2}\right) d\varepsilon \int_0^\varepsilon d\eta = \frac{1}{12} \end{aligned}$$

d'où

$$P_{y\varepsilon} = \frac{1}{12} / \sqrt{\frac{1}{4} \cdot \frac{1}{12}} = \frac{1}{\sqrt{3}}$$

Comme pour la méthode B, nous avons

$$E(Y - \varepsilon)^2 = \frac{1}{4} + \frac{1}{12} - \frac{2}{\sqrt{3}} \sqrt{\frac{1}{4} \cdot \frac{1}{12}} = \frac{1}{6}$$

Les différentes espérances mathématiques sont

$$E[F_c(\varepsilon_i)] = \int_0^1 d\varepsilon \int_0^\varepsilon d\eta = \frac{1}{2}$$

$$E[\varepsilon_i F_c(\varepsilon_i)] = \int_0^1 \varepsilon d\varepsilon \int_0^\varepsilon d\eta = \frac{1}{3}$$

Ces résultats seront utilisés dans les calculs suivants:

$$E(\varepsilon_i) = \int_0^1 x dx = \frac{1}{2}$$

$$E(\mu_i) = E[F_c(\varepsilon)([n]+1) + (1-F_c(\varepsilon))[n]] = E[F_c(\varepsilon)[n] + F_c(\varepsilon) + [n] - F_c(\varepsilon)[n]]$$

$$\begin{aligned}
 &= \mathcal{E}(F_c(\xi) + [\eta_i]) = \mathcal{E}\left(F_c(\xi) + \frac{N_i}{2} - \xi_i\right) = \frac{1}{2} + \frac{N_i}{2} - \frac{1}{2} = \frac{N_i}{2} \\
 E(\mu_i^2) &= \mathcal{E}[F_c(\xi)([\eta_i] + 1)^2 + (1 - F_c(\xi))[\eta_i]^2] = \mathcal{E}[F_c(\xi)[\eta_i]^2 + F_c(\xi_i) \\
 &\quad + 2F_c(\xi)[\eta_i] + [\eta_i]^2 - F_c(\xi)[\eta_i]] = \mathcal{E}[F_c(\xi) + 2F_c(\xi)[\eta_i] + [\eta_i]^2] \\
 &= \mathcal{E}\left[F_c(\xi) + 2F_c(\xi) \frac{N_i}{2} - 2\xi_i F_c(\xi_i) + \left(\frac{N_i}{2}\right)^2 + \xi_i^2 - 2 \frac{N_i \xi_i}{2}\right] \\
 &= \frac{1}{2} + 2 \cdot \frac{1}{2} \frac{N_i}{2} - 2 \cdot \frac{1}{3} + \left(\frac{N_i}{2}\right)^2 + \frac{1}{3} - 2 \cdot \frac{1}{2} \frac{N_i}{2} \\
 &= \left(\frac{N_i}{2}\right)^2 + \frac{1}{6}
 \end{aligned}$$

Et par conséquent

$$V(\mu_i) = E(\mu_i^2) - E(\mu_i)^2 = \frac{1}{6}$$

Pour compléter le résultat des calculs faits plus haut, nous ajoutons aux 5 méthodes étudiées, deux méthodes extrêmes hypothétiques:

La méthode extrême (1), où $p_0 = p_1 = 1/2$ et $\rho_{y\xi} = 1$

La méthode extrême (2), où $p_0 = p_1 = 1/2$ et $\rho_{y\xi} = -1$

On voit, facilement, que si ces deux méthodes existaient, elles auraient les propriétés suivantes:

	La méthode extrême (1)	La méthode extrême (2)
$V(\mu) = E(Y - \xi)^2$	$\frac{1}{12} - \left[\sqrt{\frac{1}{12}} - \frac{1}{4} \right]$	$\frac{7}{12} + \left[\sqrt{\frac{1}{12}} - \frac{1}{4} \right]$
$E(\mu)$		
$E(\mu^2)$	$\left(\frac{N}{r}\right)^2 \frac{N}{r} + \frac{1}{12} - \left[\sqrt{\frac{1}{12}} - \frac{1}{4} \right]$	$\frac{N}{r}$ $\left(\frac{N}{r}\right)^2 + \frac{7}{12} + \left[\sqrt{\frac{1}{12}} - \frac{1}{4} \right]$

Nous réunissons les propriétés de 7 méthodes dans la table suivante:

	p_1	$E(\mu)$	ρ_{yz}	$V(\mu) = E(Y - \bar{z})^2$	Efficacité par rapport à la méthode extrême (1)
Extrême	1/2	N/r	1	$\frac{1}{12} - \left[\sqrt{\frac{1}{12} - \frac{1}{4}} \right] = 0,0447$	100 %
A	1/2	N/r	$\frac{\sqrt{3}}{2}$	$1/12 = 0,0833$	53,7
A'	1/2	N/r	$-\frac{\sqrt{3}}{2}$	$7/12 = 0,5833$	7,7
A''	1/2	N/r	0	$1/3 = 0,3333$	13,4
B	1/2	N/r	$\frac{1}{\sqrt{3}}$	$1/6 = 0,1667$	25,8
C	1/2	N/r	$\frac{1}{\sqrt{3}}$	$1/6 = 0,1667$	25,8
Extrême (2)	1/2	N/r	-1	$\frac{7}{12} + \left[\sqrt{\frac{1}{12} - \frac{1}{4}} \right] = 0,6219$	7,2

Résultats:

(1) Toutes ces méthodes sont sans biais. Nous allons démontrer, au cours du § 1 de la deuxième partie, que si la méthode utilisée est sans biais, au sens de ce paragraphe et des précédents, elle l'est aussi quand nous utilisons les formules de sommation non pondérées pour l'estimation des totaux de la population.

(2) Nous avons vérifié, au cours de nos calculs, que, pour toutes les méthodes, $E(Y - \bar{z})^2 = V(\mu)$, ce que nous avons déjà démontré au commencement de ce paragraphe quand nous avons discuté à propos de la formule (15).

(3) L'efficacité de chaque méthode est une fonction croissante de la corrélation ρ_{yz} correspondante. C'est de cette

façon que nous avons pensé à l'existence, au moins hypothétique, de deux méthodes extrêmes dont les ρ_{yz} sont $+1$ et -1 . De cette façon, nous sommes assurés que toutes les méthodes sans biais possibles ont leurs $V(\mu)$ entre deux limites (hypothétiques)

$$\frac{1}{12} - \left[\sqrt{\frac{1}{12}} - \frac{1}{4} \right] \leq V(\mu) \leq \frac{7}{12} + \left[\sqrt{\frac{1}{12}} - \frac{1}{4} \right]$$

Pratiquement, nous n'avons réussi à trouver que des méthodes dont les $V(\mu)$ sont compris entre les limites

$$\frac{1}{12} \leq V(\mu) \leq \frac{7}{12}$$

Mais rien ne prouve d'ailleurs qu'on puisse trouver une méthode telle que $V(\mu) < 1/12$. Cependant c'est un problème qui reste à étudier.

(4) Nous avons calculé l'efficacité de chaque méthode par rapport à celle de la méthode extrême (1), qui serait la méthode la plus efficace si elle existait; et on voit qu'au point de vue de l'efficacité les 5 méthodes doivent être rangées dans l'ordre suivant :

La méthode A dont l'efficacité est	53,7 %
Les méthodes B et C dont les efficacités sont	25,8 %
La méthode A'' dont l'efficacité est	13,4 %
La méthode A' dont l'efficacité est	7,7 %

Afin d'adopter la meilleure méthode, on doit tenir également compte de la facilité d'emploi des méthodes, problème que nous laissons de côté.

(5) Comme ε et η sont indépendants entre eux, et des autres paramètres de la population (voir § 5 et 6), nous pouvons étendre les résultats obtenus au cours de ce paragraphe, aux sondages à plusieurs degrés, et à des combinaisons d'une ou plusieurs stratifications avec un ou plusieurs degrés de sondages, en écrivant:

$$\mu_{\alpha} = \begin{cases} [\eta_{\alpha}] + 1 & \text{avec la probabilité égale à } p_{m1} \\ [\eta_{\alpha}] & \text{avec la probabilité égale à } p_{m2} \end{cases} \quad (16)$$

L'indice α est une abréviation pour désigner une unité de sondage d'un ordre donné à tirer au sort parmi toutes celles du même ordre de notre population : par exemple, dans le cas d'une stratification suivie de deux degrés de sondages, où i indique le numéro d'ordre des strates, j celui des P S U et k celui des unités secondaires de sondage; si le plan de sondage est tel - qu'on doit tirer 1 sur r des unités secondaires de sondage, α désigné ijk , c'est-à-dire n'importe laquelle des unités secondaires.

Tout au cours de cette thèse, nous donnerons cette signification au symbole α .

m désigne A, A', A'', B ou C, suivant la méthode utilisée

$$p_{A1} = p_{A2} = p_{A'1} = p_{A'2} = p_{A''1} = p_{A''2} = 1/2$$

$$p_{B1} = \varepsilon_{\alpha}, \quad p_{B2} = 1 - \varepsilon_{\alpha}, \quad p_{C1} = F_{\alpha}(\varepsilon_{\alpha}), \quad p_{C2} = 1 - F_{\alpha}(\varepsilon_{\alpha})$$

$$E(\mu_{\alpha}) = N_{\alpha}/r \quad (17)$$

$$E(\mu_{\alpha}^2) = \left(\frac{N_{\alpha}}{r}\right)^2 + V_m \quad (m = A, A', A'', B, C) \quad (18)$$

$$\begin{cases} E(\mu_2 Z) = E(\mu) E(Z) \\ E(\mu^2 Z) = E(\mu^2) E(Z) \end{cases} \quad (19)$$

$V_m = V(\mu)$ de la méthode m utilisée; est égal à $1/12, 7/12, 1/3, 1/6$ et $1/6$ respectivement, pour les méthodes A, A', A'', B et C , Z est n'importe quelle expression contenant les paramètres de la population qui seront, suivant le cas, considérés comme des variables aléatoires.

8- Espérance mathématique de $1/\mu$

Nous avons indiqué au § 7 les limites de variation de ξ et les expressions de N en fonction de $r, [n]$ et ξ pour chaque méthode séparément. Au cours de ce paragraphe, nous allons les utiliser sans plus les mentionner.

Si nous ajoutons l'indice A, A', A'', B ou C aux symboles concernant chaque méthode ou l'indice m aux symboles concernant une méthode quelconque, nous avons :

$$\mu_m = n - V_m$$

où $n = N/r$ et

$$V_A = \begin{cases} \xi - \frac{1}{2} \\ \xi \end{cases} \quad \begin{array}{l} \text{avec la probabilité égale à } 1/2 \\ \text{" " " " } 1/2 \end{array}$$

$$V_{A'} = \begin{cases} \xi - \frac{3}{2} \\ \xi \end{cases} \quad \begin{array}{l} \text{" " " " } \\ \text{" " " " } \end{array}$$

$$V_{A''} = \begin{cases} \xi - 1 \\ \xi \end{cases} \quad \begin{array}{l} \text{" " " " } \\ \text{" " " " } \end{array}$$

$$V_B = \begin{bmatrix} \varepsilon - 1 & \text{avec la probabilité égale à} & \varepsilon \\ \varepsilon & \text{"} & \text{"} & \text{"} & 1 - \varepsilon \end{bmatrix}$$

$$V_C = \begin{bmatrix} \varepsilon_i - 1 & \text{"} & \text{"} & \text{"} & F_1(\varepsilon_i) \\ \varepsilon_i & \text{"} & \text{"} & \text{"} & 1 - F_1(\varepsilon_i) \end{bmatrix}$$

Maintenant nous calculons $E(\varepsilon^3)$ pour les 5 méthodes :

$$E_A(\varepsilon^3) = 2 \int_0^1 \frac{1}{2} \varepsilon^3 d\varepsilon = 1/2^3 (3+1)$$

$$E_{A'}(\varepsilon^3) = 2 \int_{\frac{1}{2}}^1 \varepsilon^3 d\varepsilon = [2 - (\frac{1}{2})^3] / (3+1)$$

$$E_{A'',B,C}(\varepsilon^3) = \int_0^1 \varepsilon^3 d\varepsilon$$

$$= \frac{1}{3+1}$$

Pour la méthode C nous calculons les espérances mathématiques suivantes, qui vont être utilisées un peu plus loin :

$$E[\varepsilon_i^3 F_1(\varepsilon_i)] = \int_0^1 \varepsilon_i^3 d\varepsilon_i \int_0^{\varepsilon_i} d\eta = \int_0^1 \varepsilon_i^{3+1} d\varepsilon_i = 1/(3+2)$$

$$\begin{aligned} E[(1-\varepsilon_i)^3 F_1(\varepsilon_i)] &= \int_0^1 (1-\varepsilon_i)^3 d\varepsilon_i \int_0^{\varepsilon_i} d\eta = \int_0^1 (1-\varepsilon_i)^3 \varepsilon_i d\varepsilon_i \\ &= \int_0^1 \varepsilon_i^3 d\varepsilon_i - \int_0^1 \varepsilon_i^{3+1} d\varepsilon_i = 1/(3+1)(3+2) \end{aligned}$$

Calculons alors $E(V_m^3)$ pour chaque méthode :

$$\begin{aligned} E(V_A^3) &= E\left[\frac{1}{2}(\varepsilon - \frac{1}{2})^3 + \frac{1}{2}\varepsilon^3\right] = \frac{1}{2} E(\varepsilon - \frac{1}{2})^3 + \frac{1}{2} E(\varepsilon^3) \\ &= (-1)^3 \frac{1}{2} E(\frac{1}{2} - \varepsilon)^3 + \frac{1}{2} E(\varepsilon^3) \end{aligned}$$

et utilisant la propriété de ξ d'être symétrique, nous avons

$$E(V_A^j) = (-1)^j E(\xi^j) + \frac{1}{2} E(\xi^j)$$

$$= \begin{cases} E(\xi^j) = \frac{1}{2^j(j+1)} \\ 0 \end{cases}$$

si s est pair

si s est impair

$$E(V_{A'}^j) = E\left[\frac{1}{2}\left(\xi - \frac{3}{2}\right)^j + \frac{1}{2}\xi^j\right]$$

$$= (-1)^j \frac{1}{2} E\left(\frac{3}{2} - \xi\right)^j + \frac{1}{2} E(\xi^j)$$

$$= (-1)^j \frac{1}{2} E(\xi^j) + \frac{1}{2} E(\xi^j)$$

$$= \begin{cases} E(\xi^j) = \frac{2 - \left(\frac{1}{2}\right)^j}{j+1} \\ 0 \end{cases}$$

si s est pair

si s est impair

$$E(V_{A''}^j) = E\left[\frac{1}{2}(j-1) + \frac{1}{2}\xi^j\right]$$

$$= (-1)^j \frac{1}{2} E(1-\xi)^j + \frac{1}{2} E(\xi^j)$$

$$= (-1)^j \frac{1}{2} E(\xi^j) + \frac{1}{2} E(\xi^j)$$

$$= \begin{cases} E(\xi^j) = \frac{1}{j+1} \\ 0 \end{cases}$$

Si S est pair

Si S est impair

$$\begin{aligned}
 E(V_B^s) &= E[(z-1)^s z + z^s (1-z)] \\
 &= (-1)^s E[(1-z)^s] + E[(1-z)z^s] \\
 &= (-1)^s E[z^s (1-z)] + E[(1-z)z^s] \\
 &= \begin{cases} 2 E[z^s (1-z)] = 2 E(z^s) - 2 E(z^{s+1}) = \frac{2}{(s+1)(s+2)} & \text{Si } s \text{ est pair} \\ 0 & \text{Si } s \text{ est impair} \end{cases}
 \end{aligned}$$

$$\begin{aligned}
 E(V_C^s) &= E[(z_i-1)^s F_C(z_i) + z_i^s (1-F_C(z_i))] \\
 &= (-1)^s E[(1-z_i)^s F_C(z_i)] + E(z_i^s) - E[z_i^s F_C(z_i)] \\
 &= (-1)^s \frac{1}{(s+1)(s+2)} + \frac{1}{s+1} + \frac{1}{s+2} \\
 &= \begin{cases} 2/(s+1)(s+2) & \text{Si } s \text{ est pair} \\ 0 & \text{Si } s \text{ est impair} \end{cases}
 \end{aligned}$$

On voit que, dans les 5 cas, si s est impair, $E(V_m^s) = 0$; et ici aussi les résultats des méthodes B et C sont indentiques.

Maintenant nous sommes en mesure de calculer $E(1/\mu)$ pour les 5 méthodes. Nous avons:

$$\begin{aligned}
 E\left(\frac{1}{\mu_m}\right) &= E\left(\frac{1}{n-V_m}\right) = \frac{1}{n} E\left(1 + \frac{V_m}{n} + \frac{V_m^2}{n^2} + \dots\right) \\
 &= \frac{1}{n} \left(1 + \frac{E(V_m)}{n} + \frac{E(V_m^2)}{n^2} + \dots\right)
 \end{aligned}$$

En substituant les formules obtenues plus haut dans la dernière, nous avons:

$$E\left(\frac{1}{\mu_A}\right) = \frac{1}{n} \left(1 + \frac{1}{12n^2} + \frac{1}{80n^4} + \dots\right)$$

$$E\left(\frac{1}{\mu_{A'}}\right) = \frac{1}{n} \left(1 + \frac{7}{12n^2} + \frac{15}{80n^4} + \dots\right)$$

$$E\left(\frac{1}{\mu_{A''}}\right) = \frac{1}{n} \left(1 + \frac{1}{3n^2} + \frac{1}{5n^4} + \dots\right)$$

$$E\left(\frac{1}{\mu_B}\right) = E\left(\frac{1}{\mu_C}\right) = \frac{1}{n} \left(1 + \frac{1}{6n^2} + \frac{1}{15n^4} + \dots\right)$$

Dans les 5 cas, si n est grand, les séries entre parenthèses sont rapidement convergentes. Mais comme nous avons supposé que $n > 1$, dans tous les cas elles sont convergentes; et nous pouvons garder seulement les deux premiers termes. Alors, si n est près de 2, notre erreur sera petite pour la méthode A, pas trop grande pour les méthodes B et C, mais grande pour les deux autres méthodes. Mais, en pratique, nous utiliserons la méthode A et probablement les méthodes B et C, et nous pouvons dire que pour les méthodes qui ne seront pas utilisées dans la pratique, nous gardons seulement une limite inférieure de $E(1/\mu)$. Donc, en considérant que le deuxième terme dans chaque parenthèse est égal au V_m de la méthode correspondante (voir la formule (18)), nous donnons comme formule générale pour $E(1/\mu)$:

$$\boxed{E\left(\frac{1}{\mu_m}\right) = \frac{1}{n} + \frac{V_m}{n^3}} \quad (20)$$

où $V_m = 1/12, 7/12, 1/3, 1/6$, et $1/6$, respectivement pour les méthodes A, A', A'', B et C.

9-Résumé de la première partie.

Voici, à titre de référence, les formules importantes obtenues au cours de la première partie. Ces formules seront utilisées, au cours des parties suivantes, comme bases de nos recherches.

Formules	n° des § où se trouvent les explications	n° des formules
Méthodes d'arrondissement A, A', A'', B et C	4 et 7	(1)
$N = r [n] + r$	1	(1)
$\varepsilon = a/r$	1	(1)
$\eta_i = 1 - (b_{i-1})/r$	4	
$\varepsilon = \text{rect}(0,1)$	4 et 5	
$\eta_i = \text{rect}(0,1)$	4 et 6	
$V_m = V(\mu)$	7	
$E(\mu_\alpha) = N_\alpha / r$	7	(17)
$E(\mu_\alpha^2) = (N_\alpha / r)^2 + V_m$	7	(18)
$E(\mu_z) = E(\mu)E(z)$	7	(19)
$E(\mu_z^2) = E(\mu^2)E(z)$	7	(19)
$E(1/\mu_m) = 1/n + V_m/n^3$		

V_m 1/12 , 7/12 , 1/3 , 1/6 et 1/6 , respectivement, pour les méthodes A , A' , A'' , B et C.

Deuxième partie-Les nouvelles formules d'estimateurs et de variance pour divers plans de sondage.

1- Enoncé du problème général.

La population à laquelle on va appliquer l'un des plans de sondage étudiés, possède une certaine structure: elle est découpée en unités de sondage du 1^o ordre, 2^o ordre etc.... lesquelles peuvent ou non constituer des strates.

Dans tous les cas, on se propose de soumettre à l'enquête dans chaque unité d'ordre α de sondage, $1/r_\alpha$ des unités d'ordre $\alpha+1$. (Dans le cas où on intercale entre deux degrés de sondage une stratification, nous allons remplacer " unité de degré α " dans la dernière phrase, par " strate d'ordre α "). Donc pour calculer l'effectif de sondage de l'unité n^o α , on doit diviser le nombre des unités d'ordre α ; par r_α c'est-à-dire que l'effectif de l'échantillon de l'unité n^o α est

$$n_\alpha = N_\alpha / r_\alpha \quad (1)$$

Si chaque unité de sondage contenait un nombre d'unités du degré immédiatement plus grand, multiple entier du α correspondant, les n_α seraient tous des entiers; et c'est généralement pour ce cas spécial, - ou en d'autres termes en négligeant les erreurs d'arrondissement-, qu'on établit des estimateurs sans biais de X , total des x de la population. Désignons cet estimateur par

$$\hat{X} (n) \quad (2)$$

où $n = (n, n_1, \dots, n_\alpha, \dots)$ désigne l'ensemble de tous les effectifs du plan de sondage, en considérant leurs valeurs théoriques.

Désignons la variance de l'estimateur (2) par

$$y(n) = V[X(n)] \quad (3)$$

Comme les n ne sont pas, nécessairement, des entiers, on est toujours obligé, en pratique, d'arrondir, d'une manière ou d'une autre, chaque n_x à l'un des entiers les plus proches, $[n_x]$ ou $[n_x] + 1$ que nous allons désigner par μ_x . Mais quand on veut estimer le total, on utilise le même estimateur $X(n)$ et la même formule de variance (3) en y remplaçant, par négligence, n par μ (ici et plus loin quand nous parlerons d'une façon générale, μ désigne l'ensemble $\{\mu, \mu_1, \dots, \mu_x, \dots\}$). L'estimateur qu'on utilise dans la pratique est donc

$$\hat{X}(\mu) \quad (4)$$

Nous avons donc deux problèmes:

- 1) Chercher la méthode d'arrondissement à utiliser pour que l'estimateur (4) soit sans biais, et
- 2) calculer l'augmentation de la variance causée par les erreurs d'arrondissement.

Pour résoudre ces problèmes, nous opérons pour chaque plan comme suit:

a) D'abord nous considérons un système fixe de valeurs μ , et nous calculons l'espérance mathématique E et la variance V de l'estimateur $\hat{X}(\mu)$.

a₁) Si nous avons t degrés de sondage, on a $E = E_1 E_2 \dots E_x E_t$ où chaque E_x est orthogonale aux autres.

L'estimateur $\hat{X}(n)$ est, en général, sous la forme

$$\hat{X}(n) = \sum_{i=1}^n a_i \sum_{j=1}^{n-1} b_{ij} \sum_{k=1}^{n-1} c_{ijk} \sum_{l=1}^{n-1} d_{ijkl} \dots \quad (5)$$

où S désigne la sommation limitée aux éléments de l'échantillon (tandis que nous allons toujours utiliser le symbole $\sum$ pour désigner la sommation étendue aux éléments de la population) ; $1, 2, \dots, \alpha, \dots, t$ ont les mêmes sens que celui donné plus haut à α ; les a contiennent les r_α et, probablement, d'autres paramètres de la population.

En vertu de l'orthogonalité des E_α , nous avons

$$E[\hat{X}(n)] = n \sum_1 \alpha_1 \cdot n_1 \sum_2 \alpha_2 \cdots n_{\alpha-1} \sum_\alpha \alpha_\alpha \cdots n_{t-1} \sum_t \alpha_t X_t = X \quad (6)$$

où les transformations de a en a' sont causées par les facteurs $1/N_\alpha$ dans le cas des tirages avec des probabilités égales, et par les facteurs p_α dans le cas des tirages avec des probabilités inégales p .

D'après notre hypothèse, l'estimateur (2) est sans biais. Donc dire que $\hat{X}(n)$ est sans biais est la même chose qu'accepter la formule (6).

En effectuant l'opération E sur $\hat{X}(\mu)$, nous avons

$$X^* = E[\hat{X}(\mu)] = \mu \sum_1 \alpha_1 \cdot \mu_1 \sum_2 \alpha_2 \cdots \mu_{\alpha-1} \sum_\alpha \alpha_\alpha \cdots \mu_{t-1} \sum_t \alpha_t X_t \quad (7)$$

ce qui est égal à $E[\hat{X}(n)]$ où on a remplacé n par μ

2) Afin de trouver $V[\hat{X}(\mu)]$ nous calculons

$$V[\hat{X}(\mu)] = E[\hat{X}(\mu)^2] - X^{*2}$$

On voit donc que $V[\hat{X}(\mu)]$ aussi est égale à $V[\hat{X}(n)]$ quand on y remplace n par μ (à condition que μ soit considéré comme fixe).

Nous avons donc

$$V[\hat{X}(\mu)] = y(\mu) \quad (8)$$

b_p) Nous considérons alors l'ensemble de toutes les valeurs possibles des aléatoires; et nous calculons $E[\hat{X}(\mu)] = E[E[\hat{X}(\mu)]]$ et $V[\hat{X}(\mu)]$

b₁) En considérant la formule (7), nous avons

$$\begin{aligned} E[\hat{X}(\mu)] &= E(X^*) = E(\mu) \sum_1 a'_1 \dots E(\mu_{x-1}) \sum_x a'_x \dots E(\mu_{t-1}) \sum_t a'_t x_t \\ &= n \sum_1 a'_1 \dots n_{x-1} \sum_x a'_x \dots n_{t-1} \sum_t a'_t x_t \end{aligned}$$

ce qui est égal à X d'après la formule (6); c'est-à-dire $\hat{X}(\mu)$ est un estimateur sans biais de X .

b₂)

$$\begin{aligned} V[\hat{X}(\mu)] &= E[\hat{X}(\mu) - X]^2 = E[\hat{X}(\mu) - X^* + X^* - X]^2 \\ &= E[\hat{X}(\mu) - X^*]^2 + E(X^* - X)^2 + 2E[(\hat{X}(\mu) - X^*)(X^* - X)] \\ &= E\{V[\hat{X}(\mu)]\} + V(X^*) + 2E(X^* - X)\{E[\hat{X}(\mu) - X^*]\} \end{aligned}$$

On voit immédiatement que le double produit est nul. Nous avons donc

$$V[\hat{X}(\mu)] = E\{V[\hat{X}(\mu)]\} + V(X^*) \quad (9)$$

Afin de calculer $E\{V[\hat{X}(\mu)]\}$ observons que $y(n)$ de la formule (3) est toujours sous la forme d'un polynôme comprenant chaque n , dans chaque terme; au plus sous la forme n et n^2 .

Par exemple, si nous étudions le plan formé d'une stratification suivie de deux degrés de sondage, avec des fractions de sondage $1/r$ pour les unités de premier degré et $1/s$ pour les unités de deuxième degré, nous avons

$$y(n) = r^2 s^2 \left[\sum_K \frac{n_K}{N_K} \sum_{k_i} \frac{n_{ki} (N_{ki} - n_{ki}) \sigma_{ki}^2}{N_{ki} - 1} + \sum_K \frac{n_K (N_K - n_K)}{N_K - 1} \sigma_{n_{ki}}^2 \xi_{ki} \right]$$

On voit que le premier terme de $y(n)$ comprend n_k, n_{ki} et n_{ki}^2 tandis que le deuxième terme comprend n_k, n_k^2 et $\sigma_{n_{ki}}^2 \xi_{ki}$ où n_{ki} figure au plus au deuxième degré.

Donc $y(\mu)$ aussi est un polynome contenant les μ dans chaque terme, seulement sous la forme μ et μ^2 et quand nous effectuons l'opération $\mathcal{E}'$ sur $y(\mu)$, chaque μ se transforme en n et chaque μ^2 en $n^2 + V_m$. Après avoir développé l'expression ainsi obtenue, nous aurons donc:

$$\mathcal{E}' \{ V[\hat{X}(\mu)] \} = V[\hat{X}(n)] + V_m g(V_m)$$

où g est un polynome de degré $(t-1)$ au plus en V_m et où $V_m = \sum_{i=1}^t n_i^2 - n^2$ (voir § 7 de la première partie).

De la même façon, nous pouvons démontrer que $V(\hat{X}^*)$ est égale à $V_m h(V_m)$ où h est un polynome de degré au plus $(t-1)$ en V_m .

Nous avons donc

$$V[\hat{X}(\mu)] = V_1 + V_2 \quad (10)$$

où V_1 est l'expression que nous trouverions pour $V[\hat{X}(n)]$ si nous n'avions pas tenu compte des erreurs d'arrondissement; et V_2 (qui est de la forme $V_m G(V_m)$, où G est un polynome de degré

au plus $(t-1)$ en V_m), est le résultat (ou la contribution) des erreurs d'arrondissement.

Les calculs précédents nous démontrent la proposition suivante:

Proposition- Si une méthode d'arrondissement est telle que $E(\hat{\mu}) = \mu$, c'est-à-dire si la méthode est sans biais au point de vue de la première partie, et $\hat{X}(n)$ est un estimateur sans biais de X :

1- $\hat{X}(\hat{\mu})$ est aussi un estimateur sans biais de X , et

2-La variance s'exprime sous la forme (10) ci-dessus.

Pour compléter cette proposition, nous devrions démontrer que V_2 est une fonction croissante de V_m , ce qui nous permettrait de considérer comme la meilleure, la méthode dont le V_m est le plus petit.

Nous allons démontrer cette proposition complémentaire (qui semble évidente au bon-sens), séparément, pour chacun des plans que nous voulons étudier; c'est-à-dire que nous substituerons dans les formules de nos calculs, V_m , jusqu'à ce que nous obtenions la forme explicite de V_2 ; et nous choisissons alors la meilleure méthode.

Au cours de notre deuxième partie, pour chaque plan de sondage que nous étudions, nous allons suivre les 4 étapes de calcul a_1 , a_2 , b_1 et b_2 précédentes. Nous expliquerons, seulement au passage les points qui n'ont pas été encore expliqués.

2- Liste des plans étudiés.

On peut appliquer les résultats (plus ou moins généraux) de la première partie et le § 1 de cette 2^o partie aux différents plans de sondage.

Au cours de notre partie, nous allons appliquer, à titre d'exemple, ces résultats à 6 plans de sondage. Ces plans, joints au plan que nous avons étudié dans l'introduction, constituent les 7 plans indiqués au tableau suivant.

n ^o du plan	Structure	mode de tirage	Probabilités de tirage.
I	Stratification simple (étudié au cours de l'introduction)	Exhaustif	Egales
II	Deux degrés de sondage	"	"
III	Une stratification, puis deux degrés de sondage	"	"
IV	Sondage à deux degrés avec stratification des PSU	"	"
V	2 PSU en tout	"	Inégales au 1 ^o degré
VI	Stratification, puis tirage d'une PSU de chaque strate.	Exhaustif au deuxième degré	" "
VII	Deux degrés de sondage	Bernoullien	" "

Nous nous référerons à ces plans en indiquant leur numero.

3-Plan II-Sondage à deux degrés avec probabilités égales et tirages exhaustifs.

La population contient N PSU. On se propose de soumettre à l'enquête une PSU sur r, et dans chaque PSU ainsi choisie -

un élément sur S.

L'estimateur théorique est

$$g = rs \sum_i S_i \sum_j Sx_{ij} \quad (11)$$

Lorsqu'on veut tenir compte des erreurs d'arrondissement, plusieurs sortes de problèmes peuvent se poser: le nombre N des P SU est ou n'est pas divisible par r; les nombres N_i d'éléments sont ou ne sont pas divisibles par s; enfin on peut vouloir arrondir non pas r et s, mais seulement le produit rs.

Par exemple, si $N=76$ et $N_i=100 a_i$ (a_i entier), et si je veux prendre un échantillon au centième, il n'y aurait aucune erreur d'arrondissement:

non seulement pour $r=4$, $s=25$

mais aussi pour $r=19$, $s=100/19$

On n'a pas abordé ici de problème où (rs) et r devraient être entiers à l'exclusion de s. Pour le présent plan II de sondage, on suppose que r reste égal à N/n , entier ou non; et on s'arrange pour rendre entier s (et s seulement).

Par ailleurs on a étudié plus loin (plan III) le cas où une stratification précède le sondage à deux degrés, dont le plan présent n'est qu'un cas particulier (nombre des strates égal à un); et à cette occasion sera traité le problème des erreurs d'arrondissement provenant de la double condition:

rendre r entier et s entier

Les 4 étapes de calcul introduites au § 1 sont:

$$\begin{aligned} 2_1) E(\hat{X}) &= E_1 E_2 (rS \sum_{IJ} S X_{IJ}) = rS E_1 S \cdot \frac{\mu_i}{N_i} \sum_J X_{iJ} = rS \frac{n}{N} \sum_i \frac{\mu_i}{N_i} X_i \\ &= S \sum_i \frac{\mu_i}{N_i} X_i = X^* \end{aligned}$$

$$\begin{aligned} 2_2) V(\hat{X}) &= E_1 E_2 (\hat{X} - X^*)^2 = E_1 E_2 \left(2 \sum_i S \sum_J X_{iJ} - 2 \sum_i S \mu_i \xi_{i0} \right. \\ &\quad \left. + 2 \sum_i \mu_i \xi_{i0} \right) \end{aligned}$$

où $\xi_{i0} = X_i / N_i$, et nous utiliserons ce même symbole dans tout ce qui suit sans le définir de nouveau.

Si nous mettons

$$\bar{x}_{i0} = \frac{S_j X_{iJ}}{\mu_i}$$

$$\overline{\mu_i \xi_{i0}} = \frac{S_i \mu_i \xi_{i0}}{n} \quad \text{et} \quad \overline{\mu_i \xi_{i0}} = \frac{\sum \mu_i \xi_{i0}}{N}$$

nous avons

$$\begin{aligned} V(\hat{X}) &= E_1 E_2 \left[2 \sum_i \mu_i (\bar{x}_{i0} - \xi_{i0}) + N (\overline{\mu_i \xi_{i0}} - \overline{\mu_i \xi_{i0}}) \right]^2 \\ &= E_1 \left[2 \sum_i \mu_i^2 \bar{x}_{i0}^2 + N^2 (\overline{\mu_i \xi_{i0}} - \overline{\mu_i \xi_{i0}})^2 \right] \\ &\quad + E_1 2 \sum_{i \neq j} S_i S_j \mu_i \mu_j E_2 (\bar{x}_{i0} - \xi_{i0})(\bar{x}_{j0} - \xi_{j0}) \\ &\quad + 2 E_1 \left[N (\overline{\mu_i \xi_{i0}} - \overline{\mu_i \xi_{i0}}) \sum_i \mu_i E_2 (\bar{x}_{i0} - \xi_{i0}) \right] \end{aligned}$$

Comme nos tirages sont exhaustifs, dans chacun des doubles produits les deux facteurs qui se multiplient sont différents (tandis que, si nos tirages étaient bernoulliens, il serait possible que dans quelques-une des produits les deux facteurs

soient identiques). Nous avons donc

$$E_2(\bar{x}_{i0} - \bar{x}_{i0})(\bar{x}_{i0} - \bar{x}_{i0}) = E_2(\bar{x}_{i0} - \bar{x}_{i0}) E_2(\bar{x}_{i0} - \bar{x}_{i0})$$

et par conséquent, les doubles produits sont nuls; et nous avons

$$\begin{aligned} V(\hat{X}) &= \sigma^2 E_1 \left[\sum_i \frac{N_i \mu_i^2 \sigma_i^2}{(N_i - 1) \mu_i} + N (\mu_i \bar{x}_{i0} - \mu_i \bar{x}_{i0}) \right] \\ &= \sigma^2 \left[\frac{n}{N} \sum_i \frac{N_i \mu_i^2 - \mu_i^2}{N_i - 1} \sigma_i^2 + \frac{N(N-n)}{N-1} \frac{\sigma_{\mu_i \bar{x}_{i0}}^2}{n} \right] \\ &= \sigma^2 \left[\sum_i \frac{N \mu_i - \mu_i^2}{N_i - 1} \sigma_i^2 + \frac{N(n-1)}{N-1} \sigma_{\mu_i \bar{x}_{i0}}^2 \right] \end{aligned}$$

b₁)

$$\begin{aligned} E(\hat{X}) &= E[E(\hat{X})] = E(X^*) = E \left[\sum_i \frac{\mu_i}{N_i} X_i \right] \\ &= \sum_i \frac{X_i}{N_i} E(\mu_i) = \sum_i \frac{X_i}{N_i} \frac{N_i}{\sigma} = \sum_i X_i = X \end{aligned}$$

b₂) Calcul de $V(\hat{X})$

$$\begin{aligned} V(X^*) &= \sigma^2 \sum_i \left(\frac{X_i}{N_i} \right)^2 V(\mu_i) = V_{\mu_i} \sigma^2 \sum_i \frac{X_i^2}{N_i} \\ E V(\hat{X}) &= \sigma^2 \left[\sum_i \frac{N_i E(\mu_i^2) - E(\mu_i^2)}{N_i - 1} \sigma_i^2 + \frac{N(n-1)}{N-1} E(\sigma_{\mu_i \bar{x}_{i0}}^2) \right] \end{aligned}$$

Maintenant nous calculons $E(\sigma_{\mu_i \bar{x}_{i0}}^2)$

$$\begin{aligned} E(\sigma_{\mu_i \bar{x}_{i0}}^2) &= E \left\{ \frac{1}{N} \left[\sum_i \mu_i^2 \bar{x}_{i0}^2 - \frac{1}{N} \left(\sum_i \mu_i \bar{x}_{i0} \right)^2 \right] \right\} \\ &= E \left\{ \frac{1}{N} \left[\sum_i \mu_i^2 \bar{x}_{i0}^2 - \frac{1}{N} \sum_i \mu_i^2 \bar{x}_{i0}^2 - \frac{1}{N} \sum_{i \neq i'} \mu_i \mu_{i'} \bar{x}_{i0} \bar{x}_{i0} \right] \right\} \\ &= \frac{1}{N} \left[\sum_i \bar{x}_{i0}^2 E(\mu_i^2) \left(1 - \frac{1}{N} \right) - \frac{1}{N} \sum_{i \neq i'} \sum_i \bar{x}_{i0} \bar{x}_{i0} E(\mu_i \mu_{i'}) \right] \end{aligned}$$

Comme le μ de chaque PSU est indépendant de ceux des autres PSU, on peut remplacer $E(\mu_i \mu_{i'})$ par $E(\mu_i)E(\mu_{i'})$; et nous avons

$$E(\sigma_{\mu_i \xi_{i0}}^2) = \frac{1}{N} \left[\left(1 - \frac{1}{N}\right) \sum_i \xi_{i0}^2 \left(\frac{N_i}{j^2} + V_m \right) - \frac{1}{N} \sum_i \sum_{i' \neq i} \xi_{i0} \xi_{i'0} \frac{N_i}{j} \frac{N_{i'}}{j} \right]$$

où nous avons utilisé les formules (17) et (18) de la première partie.

D'où

$$\begin{aligned} E(\sigma_{\mu_i \xi_{i0}}^2) &= \frac{1}{N} \left[\left(1 - \frac{1}{N}\right) \left(\frac{1}{j^2} \sum_i X_i^2 + V_m \sum_i \xi_{i0}^2 \right) - \frac{1}{N j^2} \sum_i \sum_{i' \neq i} X_i X_{i'} \right] \\ &= \frac{1}{N} \left[\left(1 - \frac{1}{N}\right) \frac{1}{j^2} \sum_i X_i^2 + \left(1 - \frac{1}{N}\right) V_m \sum_i \xi_{i0}^2 - \frac{1}{N j^2} (X - \sum_i X_i^2) \right] \\ &= \frac{1}{N} \left[\frac{1}{j^2} \sum_i X_i^2 + \left(1 - \frac{1}{N}\right) V_m \sum_i \xi_{i0}^2 - \frac{X^2}{N j^2} \right] \end{aligned}$$

Et en posant

$$\sigma_X^2 = \frac{1}{N} \left(\sum_i X_i^2 - \frac{X^2}{N} \right)$$

nous avons

$$E(\sigma_{\mu_i \xi_{i0}}^2) = \frac{1}{j^2} \sigma_X^2 + \left(\frac{1}{N} - \frac{1}{N j^2} \right) V_m \sum_i \xi_{i0}^2$$

Nous avons donc

$$\begin{aligned} V(\hat{X}) &= j^2 \sum_i \frac{\frac{N_i^2}{j^2} - \left(\frac{N_i}{j}\right)^2 - V_m \sigma_i^2}{N_i - 1} \\ &\quad + \frac{N^2 (j-1)}{N-1} \sigma_X^2 + j^2 V_m \sum_i \xi_{i0}^2 \\ &\quad + \frac{V_m j^2 N^2 (j-1)}{N-1} \left(\frac{1}{N} - \frac{1}{N j^2} \right) \sum_i \xi_{i0}^2 \end{aligned}$$

$$= V_1 + V_2 \quad (12)$$

où

$$V_1 = z(z-1) \sum_c \frac{N_c^2 \sigma_c^2}{N_c - 1} + \frac{N(z-1)}{N-1} \sigma_X^2$$

$$V_2 = z^2 V_m \left[\sum_c \xi_{c0}^2 - \sum_c \frac{\sigma_c^2}{N_c - 1} \right]$$

On voit que V_2 est proportionnelle à V_m , ce qui vérifie la proposition donnée, intuitivement, à la fin du §1. On voit donc que dans le cas présent, la meilleure méthode est la méthode A (avec $V_m = 1/12$)

4- Plan III-Stratification suivie de deux degrés de sondage, avec probabilités égales et tirages exhaustifs.

La population contient N strates dont le $n^{\circ}k$ contient N_k PSU; et la PSU $n^{\circ}ki$ renferme N_{ki} éléments. On se propose de soumettre à l'enquête $1/r$ des PSU dans chaque strate, et $1/s$ des éléments dans chaque PSU désignée, par tirages exhaustifs et avec probabilités égales.

L'estimateur théorique est

$$\hat{X} = z s \sum_K S_c \sum_J x_{kij} \quad (13)$$

a₁)

$$\begin{aligned} E(\hat{X}) &= E_1 E_2 \left[z s \sum_K S_c \sum_J x_{kij} \right] = z s E_1 \sum_K S_c \frac{\mu_{ki}}{N_{ki}} \sum_J x_{kij} \\ &= z s E_1 \sum_K S_c \frac{\mu_{ki}}{N_{ki}} X_{ki} = z s \sum_K \frac{\mu_K}{N_K} \sum_c \frac{\mu_{ki}}{N_{ki}} X_{ki} = X^* \end{aligned}$$

a₂)

$$V(\hat{X}) = E(\bar{X} - X^*)^2 = E_1 E_2 \left(2\beta \sum_K \sum_{i,j} \delta_{ij} x_{Kij} - 2\beta \sum_K \sum_i \mu_{Ki} \xi_{Ki} + 3\beta \sum_K \sum_i \mu_{Ki} \xi_{Kio} - 2\beta \sum_K \frac{\mu_K}{N_K} \sum_i \mu_{Ki} \xi_{Kio} \right)^2$$

où nous avons ajouté et retranché un même terme, et $\xi_{Kio} = \sum_j x_{Kij} / N_K$
En posant

$$\frac{1}{\mu_{Ki}} \sum_j x_{Kij} = \bar{x}_{Kio}, \quad \frac{1}{\mu_K} \sum_i \mu_{Ki} \xi_{Kio} = \overline{\mu_{Ki} \xi_{Kio}}$$

et $\frac{1}{N_K} \sum_i \mu_{Ki} \xi_{Kio} = \overline{\mu_{Ki} \xi_{Kio}}$ nous avons

$$V(\hat{X}) = 2\beta^2 E_1 E_2 \left[\sum_K \sum_i \mu_{Ki} (\bar{x}_{Kio} - \xi_{Kio})^2 + \sum_K \mu_K (\overline{\mu_{Ki} \xi_{Kio}} - \overline{\mu_K \xi_{Kio}})^2 \right]^2$$

$$= 2\beta^2 \left[\sum_K \sum_i \mu_{Ki} E_1 E_2 (\bar{x}_{Kio} - \xi_{Kio})^2 + \sum_K \mu_K^2 E_1 (\overline{\mu_{Ki} \xi_{Kio}} - \overline{\mu_K \xi_{Kio}})^2 \right]$$

car pour les mêmes raisons déjà données au §3, les doubles produits sont nuls. Il vient

$$V(\hat{X}) = 2\beta^2 \left[\sum_K \frac{\mu_K}{N_K} \sum_i \frac{\mu_{Ki} (N_{Ki} - \mu_{Ki})}{N_{Ki} - 1} \frac{\sigma_{Ki}^2}{\mu_{Ki}} + \sum_K \frac{\mu_K (N_K - \mu_K)}{N_K - 1} \frac{\sigma_{\mu_{Ki} \xi_{Kio}}^2}{\mu_K} \right]$$

$$= \sum_K \frac{\mu_K^2}{N_K} \sum_L \frac{\mu_{KL} (N_{KL} - \mu_{KL})}{N_{KL} - 1} \sigma_{KL}^2 + \sum_K \frac{\mu_K (N_K - \mu_K)}{N_K - 1} \sigma_{\mu_{KL}}^2$$

$$b_1) \quad E(\hat{X}) = E[E(\hat{X})] = E_1 E_2 [E(\hat{X})] = \sum_K \frac{E(\mu_K)}{N_K} \sum_L \frac{E(\mu_{KL})}{N_{KL}} X_{KL} \\ = \sum_K \frac{N_K}{N_K} \sum_L \frac{N_{KL}}{N_{KL}} X_{KL} = \sum_K \sum_L X_{KL} = X$$

$$b_2) \quad V(\hat{X}) = E_1 E_2 \left(\sum_K \frac{\mu_K}{N_K} \sum_L \frac{\mu_{KL}}{N_{KL}} X_{KL} - \sum_K \frac{\mu_K}{N_K} \sum_L \frac{1}{s} X_{KL} \right. \\ \left. + \sum_K \frac{\mu_K}{N_K} X_K - \sum_K X_K \right)^2$$

où $X_K = \sum_L X_{KL}$, et où nous avons (une fois) ajouté et retranché un même terme dans la parenthèse, car $\sum_L X_{KL}/s = X_K$. Donc

$$V(X^*) = E_1 E_2 \left[\sum_K \frac{\mu_K}{N_K} \sum_L \frac{X_{KL}}{N_{KL}} \left(\mu_{KL} - \frac{N_{KL}}{s} \right) \right. \\ \left. + \sum_K \frac{X_K}{N_K} \left(\mu_K - \frac{N_K}{s} \right) \right]^2$$

et en posant $X_K/N_K = \bar{s}_{K0}$ nous avons

$$V(X^*) = \sum_K \frac{E_1(\mu_K^2)}{N_K^2} \sum_L \bar{s}_{KL0} E_2 \left(\mu_{KL} - \frac{N_{KL}}{s} \right)^2 \\ + \sum_K \bar{s}_{K0}^2 E_1 \left(\mu_K - \frac{N_K}{s} \right)^2 + E[\text{double products}]$$

Comme les μ_K et μ_{KL} sont indépendants et nos tirages exhaustifs, les doubles produits sont nuls; et en tenant compte des formules (17) et (18) de la première partie, nous avons:

$$\begin{aligned}
 V(X^*) &= \sum_K^2 \sum_{K_0}^2 \frac{E(M_K)^2}{N_K^2} \sum_{K_i}^2 \xi_{K_i 0}^2 V(M_{K_i}) \\
 &+ \sum_K^2 \sum_{K_0}^2 \xi_{K_0}^2 V(M_K) = \sum_K^2 \sum_{K_0}^2 \frac{(M_K)^2}{N_K^2} + V_m \sum_{K_i}^2 \xi_{K_i 0}^2 V_m \\
 &+ \sum_K^2 \sum_{K_0}^2 \xi_{K_0}^2 V_m \\
 &= V_m \sum_K^2 \sum_{K_i}^2 \xi_{K_i 0}^2 + \sum_K^2 V_m \sum_{K_0}^2 \xi_{K_0}^2 \\
 &+ \sum_K^2 \sum_{K_0}^2 V_m \sum_{K_i}^2 \frac{1}{N_K} \sum_{K_i}^2 \xi_{K_i 0}^2
 \end{aligned}$$

D' autre part, nous avons

$$\begin{aligned}
 E'[V(\hat{X})] &= \sum_K^2 \sum_{K_0}^2 E'_1 E'_2 \left[\sum_K^2 \frac{M_K}{N_K} \sum_{K_i}^2 \frac{M_{K_i} (N_{K_i} - M_{K_i})}{N_{K_i} - 1} \sigma_{K_i 0}^2 \right. \\
 &\quad \left. + \sum_K^2 \frac{M_K (N_K - M_K)}{N_K - 1} \sigma_{M_{K_i} \xi_{K_i 0}}^2 \right] \\
 &= \sum_K^2 \sum_{K_0}^2 \left[\sum_K^2 \frac{E(M_K)}{N_K} \sum_{K_i}^2 \frac{N_{K_i} E(M_{K_i}) - E(M_{K_i}^2)}{N_{K_i} - 1} \sigma_{K_i}^2 \right. \\
 &\quad \left. + \sum_K^2 \frac{N_K E(M_K) - E(M_K^2)}{N_K - 1} E(\sigma_{M_{K_i} \xi_{K_i 0}}^2) \right]
 \end{aligned}$$

Notons que $\sigma_{K_i \xi_{K_i 0}}^2$ dépend, seulement, des M_{K_i} et non pas des M_K . C'est pourquoi l'espérance mathématique du deuxième terme de l'accolade (de la forme $\sum_K A_K B_K$) a pu être prise égale à

$$\sum_K E(A_K) E(B_K)$$

$$E[V(\hat{X})] = \frac{1}{2} \sum_K \frac{1}{N_K} \sum_i \frac{\frac{N_{Ki}^2}{2} - (\frac{N_{Ki}}{2})^2 - V_m}{N_{Ki} - 1} \sigma_{Ki}^2 + \sum_K \frac{\frac{N_K^2}{2} - (\frac{N_K}{2})^2 - V_m}{N_K - 1} E(\sigma_{\mu_{Ki}}^2 \xi_{\mu_{Ki}0})]$$

Pour calculer $\sigma_{\mu_{Ki}}^2 \xi_{\mu_{Ki}0}$ nous pouvons utiliser le même procédé qu'au § 3 pour calculer $\sigma_{\mu_i}^2 \xi_{\mu_i0}$ avec cette différence que chaque i doit être précédé par k , ce qui donne:

$$E(\sigma_{\mu_{Ki}}^2 \xi_{\mu_{Ki}0}) = \frac{1}{2} \sigma_{XK}^2 + \left(\frac{1}{N_K} - \frac{1}{N_K^2} \right) V_m \sum_i \xi_{KLi}^2$$

où $\sigma_{XK}^2 = \frac{1}{N_K} \left(\sum_i X_{Ki}^2 - \frac{X_K^2}{N_K} \right)$

Donc

$$\begin{aligned} V(\hat{X}) = & V_m \sum_K \sum_i \xi_{KLi}^2 + \frac{1}{2} V_m \sum_K \frac{1}{N_K} \sum_i \xi_{KLi}^2 \\ & + V_m \sum_K \xi_{K0}^2 + \frac{1}{2} \sum_K \frac{(2-1)}{2} \sum_i \frac{N_{Ki}}{N_{Ki}-1} \sigma_{Ki}^2 \\ & - \frac{1}{2} V_m \sum_K \sum_i \frac{\sigma_{Ki}^2}{N_{Ki}-1} \\ & + \frac{1}{2} \sum_K \frac{N_K^2 (2-1)}{2^2 (N_K-1)} \frac{\sigma_{XK}^2}{2} \\ & + \frac{1}{2} V_m \sum_K \frac{N_K^2 (2-1)}{2^2 (N_K-1)} \left(\frac{1}{N_K} - \frac{1}{N_K^2} \right) \sum_i \xi_{K0}^2 \\ & - \frac{1}{2} V_m \sum_K \frac{\sigma_{XK}^2}{N_K-1} - V_m \sum_K \frac{1}{N_K} \sum_i \xi_{KLi}^2 \end{aligned}$$

ce qui donne:

$$V(\hat{X}) = V_1 + V_2 \quad (14)$$

avec

$$\begin{aligned} V_1 &= z(z-1) \sum_K \sum_L \frac{N_{KL} \sigma_{KL}^2}{N_K - 1} + (z-1) \sum_K \frac{N_K \sigma_{xK}^2}{N_K - 1} \\ V_2 &= V_m z^2 \left[\sum_K \sum_L \frac{\sigma_{KL}^2}{N_{KL} - 1} \right] \\ &\quad + V_m z^2 \left[\sum_K \frac{\sigma_{xK}^2}{N_K - 1} \right] \end{aligned}$$

5- Plan IV-Sondage à deux degrés avec stratification des PSU.

Il est possible que la population soit divisée trois fois de suite c'est-à-dire la population en N PSU, et chaque PSU en N_1 unités secondaires de sondage, dont la $n^{o}ij$ contient elle-même N_{ij} éléments. Par exemple on peut prendre les districts ruraux comme PSU, chaque village comme unité d'ordre 2, et chaque paysan comme élément. Quelquefois (comme dans l'exemple mentionné, la variance entre les PSU (ici les variations géographiques) est plus petite que la variance entre les unités du deuxième ordre (ici entre les villages). On se propose, en pareil cas, de stratifier l'intérieur des PSU, c'est-à-dire de soumettre à l'enquête $1/r$ (par exemple) des PSU, puis dans celles-ci, toutes les unités du 2ième ordre considérées comme des strates, et enfin $1/s_{ij}$ des éléments de la strate $n^{o}ij$ (où i désigne une PSU choisie). Nous allons étudier cet exemple en supposant les probabilités égales et les tirages exhaustifs.

L'estimateur théorique ici est

$$\hat{X} = z \sum_i \sum_j s_{ij} \sum_K x_{iJK}$$

Ici aussi, pour simplifier les calculs, nous adoptons la même hypothèse que dans le cas du sondage à deux degrés (plan II), c'est-à-dire que nous supposons que r (entier ou non) ne produit pas d'erreur d'arrondissement.

Les 4 étapes de calcul sont:

a1)

$$\begin{aligned} E(\hat{X}) &= E_1 E_2(\hat{X}) = E_1 \left[z \sum_i \sum_j \frac{\mu_{ij}}{N_{ij}} X_{ij} \right] = z \frac{1}{z} \sum_i \sum_j s_{ij} \frac{\mu_{ij}}{N_{ij}} X_{ij} \\ &= \sum_i \sum_j s_{ij} \frac{\mu_{ij}}{N_{ij}} X_{ij} = \sum_i Z_i = X^* \\ \text{ou } Z_i &= \sum_j s_{ij} \frac{\mu_{ij}}{N_{ij}} X_{ij} \end{aligned}$$

a2)

$$\begin{aligned} V(\hat{X}) &= E(\hat{X} - X^*)^2 = E \left[z \sum_i \sum_j s_{ij} \sum_K x_{iJK} - X^* \right]^2 \\ &= E \left[z \sum_i \sum_j s_{ij} \sum_K x_{iJK} - z \sum_i \sum_j s_{ij} \frac{\mu_{ij}}{N_{ij}} \sum_K x_{iJK} \right. \\ &\quad \left. + z \sum_i \sum_j s_{ij} \frac{\mu_{ij}}{N_{ij}} \sum_K x_{iJK} - X^* \right]^2 \\ &= E \left[z \sum_i \sum_j s_{ij} \mu_{ij} (\bar{x}_{ij0} - \bar{s}_{ij0}) + N(\bar{z} - \bar{X}) \right]^2 \\ \text{ou } \bar{x}_{ij0} &= \frac{1}{\mu_{ij}} \sum_K x_{iJK} \end{aligned}$$

$$\xi_{ij0} = \frac{1}{N_{ij}} \sum_K X_{iJK}$$

$$\bar{Z} = \frac{1}{N} \sum_i Z_i \quad \bar{\bar{Z}} = \frac{1}{N} \sum_i \bar{Z}_i$$

Nous avons donc

$$V(\hat{X}) = E_1 \left[\sum_i \sum_j \delta_{ij}^2 \mu_{ij}^2 \frac{N_{ij} - \mu_{ij}}{N_{ij} - 1} \frac{\sigma_{ij}^2}{\mu_{ij}} + N^2 (\bar{Z} - \bar{\bar{Z}})^2 \right]$$

$$+ E_2 (\text{double Products})$$

Comme E_2 (doubles produits) est (pour les mêmes raisons que nous avons données au cours du § 3) égal à zéro,

$$\begin{aligned} V(\hat{X}) &= E_1 \frac{1}{N} \sum_i \sum_j \delta_{ij}^2 \frac{\mu_{ij}(N_{ij} - \mu_{ij})}{N_{ij} - 1} \sigma_{ij}^2 + \frac{N(N - \frac{N}{2})}{N - 1} \frac{\sigma_Z^2}{N/2} \\ &= E_1 \sum_i \sum_j \delta_{ij}^2 \frac{\mu_{ij}(N_{ij} - \mu_{ij})}{N_{ij} - 1} \sigma_{ij}^2 + \frac{N(2-1)}{N-1} \sigma_Z^2 \end{aligned}$$

b1)

$$\begin{aligned} E(\hat{X}) &= E_1 E_2(\hat{X}) = E_1(X^*) = \sum_i \sum_j \delta_{ij} \frac{E(\mu_{ij})}{N_{ij}} X_{ij} \\ &= \sum_i \sum_j \frac{N_{ij} \delta_{ij}}{N_{ij}} X_{ij} = \sum_i \sum_j X_{ij} = X \end{aligned}$$

b2)

$$E[V(\hat{X})] = E_1 \sum_i \sum_j \frac{N_{ij} E(\mu_{ij}) - E(\mu_{ij}^2) \sigma_{ij}^2}{N_{ij} - 1} + \frac{N(2-1)}{N-1} E(\sigma_Z^2)$$

$$E(\sigma_Z^2) = E_1 \frac{1}{N} \left[\sum_i \left(\sum_j \frac{\delta_{ij} \mu_{ij}}{N_{ij}} X_{ij} \right)^2 - \frac{1}{N} \left(\sum_i \sum_j \frac{\delta_{ij} \mu_{ij}}{N_{ij}} X_{ij} \right)^2 \right]$$

$$\begin{aligned}
 &= E' \frac{1}{N} \left\{ \left(1 - \frac{1}{N}\right) \left[\sum_c \sum_j \left(\frac{s_{ij} \mu_{ij}}{N_{ij}} \right) X_{ij}^2 + \sum_c \sum_{j \neq j'} \sum_{j'} \frac{s_{ij} s_{ij'} \mu_{ij} \mu_{ij'}}{N_{ij} N_{ij'}} X_{ij} X_{ij'} \right] \right. \\
 &\quad \left. \left(1 - \frac{1}{N}\right) - \frac{1}{N} \sum_{i \neq c'} \sum_j \sum_{j'} \frac{s_{ij} s_{ij'} \mu_{ij} \mu_{ij'}}{N_{ij} N_{ij'}} X_{ij} X_{ij'} \right\} \\
 &= \frac{1}{N} \left\{ \left(1 - \frac{1}{N}\right) \left[\sum_c \sum_j s_{ij} \left(\frac{N_{ij}^2}{s_{ij}^2} + V_m \right) \left(\frac{X_{ij}}{N_{ij}} \right)^2 \right] \right. \\
 &\quad \left. - \frac{1}{N} \sum_{i \neq c'} \sum_j \sum_{j'} \frac{s_{ij} s_{ij'} \left(\frac{N_{ij}}{s_{ij}} \right) \left(\frac{N_{ij'}}{s_{ij'}} \right) X_{ij} X_{ij'}}{s_{ij} s_{ij'}} \right\} \\
 &\quad + \sum_c \sum_{j \neq j'} \sum_{j'} \frac{s_{ij} s_{ij'} (N_{ij}/s_{ij}) (N_{ij'}/s_{ij'}) X_{ij} X_{ij'}}{N_{ij} N_{ij'}}
 \end{aligned}$$

(où nous avons utilisé les formules (19), et (17); et (18) de la lière partie).

$$\begin{aligned}
 &= \frac{1}{N} \left\{ \left(1 - \frac{1}{N}\right) \left[\sum_c \sum_j \left(1 + \frac{s_{ij}^2 V_m}{N_{ij}^2} \right) X_{ij}^2 + \sum_c \sum_{j \neq j'} \sum_{j'} X_{ij} X_{ij'} \right] \right. \\
 &\quad \left. - \frac{1}{N} \sum_{i \neq c'} \sum_j \sum_{j'} X_{ij} X_{ij'} \right\} \\
 &= \frac{1}{N} \left\{ \left(1 - \frac{1}{N}\right) \sum_c X_c^2 + \left(1 - \frac{1}{N}\right) V_m \sum_c \sum_j \frac{s_{ij}^2}{N_{ij}^2} X_{ij}^2 - \frac{1}{N} \sum_{i \neq c'} \sum_{j'} X_{ij}^2 \right\} \\
 &= \frac{1}{N} \left\{ \left(1 - \frac{1}{N}\right) V_m \sum_c \sum_j \frac{s_{ij}^2}{N_{ij}^2} X_{ij}^2 + \left(1 - \frac{1}{N}\right) \sum_c X_c^2 - \frac{1}{N} (X^2 - \sum_c X_c^2) \right\} \\
 &= \frac{1}{N} \left\{ \sum_c X_c^2 + \left(1 - \frac{1}{N}\right) V_m \sum_c \sum_j \frac{s_{ij}^2}{N_{ij}^2} X_{ij}^2 - \frac{X^2}{N} \right\} \\
 &= \sigma_X^2 + \left(\frac{1}{N} - \frac{1}{N^2} \right) V_m \sum_c \sum_j s_{ij}^2 s_{ij}^2 \\
 &\quad \text{où} \quad \sigma_X^2 = \frac{1}{N} \left[\sum_c X_c^2 - \frac{X^2}{N} \right]
 \end{aligned}$$

Nous avons donc

$$\begin{aligned} E[V(\hat{X})] &= E \sum_i \sum_j \delta_{ij}^2 \frac{\frac{N_{ij}^2}{\delta_{ij}^2} - \left(\frac{N_{ij}}{\delta_{ij}}\right)^2 - V_m}{N_{ij} - 1} \sigma_{ij}^2 \\ &\quad + \frac{N^2(z-1)}{N-1} \left[\sigma_x^2 + \left(\frac{1}{N} - \frac{1}{N^3}\right) V_m \sum_i \sum_j \delta_{ij}^2 \xi_{ij}^2 \right] \\ &= E \sum_i \sum_j \frac{N_{ij}(\delta_{ij} - 1)}{N_{ij} - 1} \sigma_{ij}^2 \\ &\quad + N \frac{z(z-1)}{N-1} \sigma_x^2 - E V_m \sum_i \sum_j \frac{\delta_{ij}^2 \sigma_{ij}^2}{N_{ij} - 1} \\ &\quad + (z-1) V_m \sum_i \sum_j \delta_{ij}^2 \xi_{ij}^2 \end{aligned}$$

$$\begin{aligned} V(X^*) &= \sum_i \sum_j \delta_{ij}^2 \frac{V(\delta_{ij}^2)}{N_{ij}^2} = \sum_i \sum_j \delta_{ij}^2 \frac{V_m}{N_{ij}^2} X_{ij}^2 \\ &= V_m \sum_i \sum_j \delta_{ij}^2 \xi_{ij}^2 \end{aligned}$$

Donc

$$\begin{aligned} V(\hat{X}) &= E \sum_i \sum_j \frac{N_{ij}(\delta_{ij} - 1)}{N_{ij} - 1} \sigma_{ij}^2 \\ &\quad + \frac{N(z-1)\sigma_x^2}{N-1} + (z-1) V_m \sum_i \sum_j \delta_{ij}^2 \xi_{ij}^2 \\ &\quad - E V_m \sum_i \sum_j \frac{\delta_{ij}^2 \sigma_{ij}^2}{N_{ij} - 1} \\ &\quad + V_m \sum_i \sum_j \delta_{ij}^2 \xi_{ij}^2 \\ &= V_1 + V_2 \quad (15) \end{aligned}$$

$$\begin{aligned} \hat{V}_1 &= r \sum_i \sum_j \frac{N_{ij}^2 (s_{ij} - 1) \sigma_{ij}^2}{N_{ij} - 1} + \frac{N^2 (r-1) \sigma_{\epsilon}^2}{N-1} \\ \hat{V}_2 &= r V_m \left[\sum_i \sum_j s_{ij}^2 s_{ij}^2 - \sum_i \sum_j \frac{s_{ij}^2 \sigma_{ij}^2}{N_{ij} - 1} \right] \end{aligned}$$

On voit qu'ici aussi V_2 est proportionnelle à V_m et que, par conséquent, la méthode A (avec $V_m = 1/12$) est la meilleure.

6- Plan V-Sondage de deux PSU avec probabilités inégales et tirages exhaustifs.

La population possède N PSU, la PSU n°i contenant N_i éléments. On fait correspondre à chaque PSU une probabilité de tirage telle que $p_i \neq p_j$. On se propose de choisir la PSU n°i avec la probabilité p_i , et de cette façon on choisit $n=2$ PSU tirage exhaustif; et on s'arrange pour qu'en moyenne, de chaque PSU ainsi choisie on tire un élément sur r .

Appelons P_i la probabilité totale pour que la PSU n°i fasse partie de l'échantillon, et P_{ij} celle pour que les PSU n°i et n°j soient toutes deux, dans l'échantillon.

On voit, facilement, que l'effectif de l'échantillon dans la PSU n°i doit être

$$n_i = N_i / Nr P_i \quad (16)$$

car, dans ce cas, nous avons

$$E(n_i) = \sum_i P_i \frac{N_i}{Nr P_i} = \frac{\sum N_i}{Nr} = \frac{N_0}{Nr} = \frac{\bar{N}_0}{r}$$

d'où

$$r = \bar{N}_0 / E(n_i)$$

Un estimateur sans biais de X, total de la population, est

$$\hat{X} = N^2 \sum_i^n \sum_j^{H_i} x_{ij} \quad (17)$$

Les 4 étapes de calcul sont:

a1)

$$\begin{aligned} E(\hat{X}) &= N^2 \sum_1^n E_1 \sum_2^{H_i} x_{ij} \\ &= N^2 \sum_1^n E_1 \sum_2^{H_i} x_{ij} = N^2 \sum_1^n E_1 \sum_2^{H_i} \frac{H_i}{N_i} x_{ij} \\ &= N^2 \sum_1^n P_i \frac{H_i}{N_i} X_i = X^* \end{aligned}$$

a2)

$$\begin{aligned} V(\hat{X}) &= E_1 E_2 \left[N^2 \sum_i^n \sum_j^{H_i} x_{ij} - N^2 \sum_i^n H_i \xi_{i0} \right. \\ &\quad \left. + N^2 \sum_i^n H_i \xi_{i0} + N^2 \sum_i^n P_i \frac{H_i}{N_i} X_i \right]^2 \\ &= N^2 E_1 E_2 \left[\sum_i^n H_i (\bar{x}_{i0} - \xi_{i0}) + \left\{ \sum_i^n H_i \xi_{i0} - \left(\sum_i^n H_i \xi_{i0} \right) \right\} \right]^2 \\ &= N^2 E_1 \sum_i^n H_i E_2 (\bar{x}_{i0} - \xi_{i0})^2 + N^2 E_1 \sum_i^n \sum_{i \neq i'}^{H_i} H_i H_{i'} \\ &\quad + E_2 (\bar{x}_{i0} - \xi_{i0})(\bar{x}_{i'0} - \xi_{i'0}) + N^2 E_1 \left[\sum_i^n H_i \xi_{i0} - E(\sum_i^n H_i \xi_{i0}) \right]^2 \\ &\quad + 2 E_1 \left[\sum_i^n H_i \xi_{i0} - E(\sum_i^n H_i \xi_{i0}) \right] \left\{ E_2 \sum_i^n H_i (\bar{x}_{i0} - \xi_{i0}) \right\} \end{aligned}$$

Avec les mêmes raisonnements déjà donnés pour les plans précédents, nous pouvons démontrer que les doubles-produits sont nuls, et nous avons:

$$\begin{aligned}
 V(\hat{X}) &= N^2 \sum_i \sum_j P_i P_j \frac{2N_i - M_i}{N_i - 1} \frac{\sigma_i^2}{M_i} + N^2 \left[\left(\sum_i M_i \xi_{i0} \right)^2 + \left(\sum_i P_i M_i \xi_{i0} \right)^2 \right. \\
 &\quad \left. - 2 \left(\sum_i U_i \xi_{i0} \right) \left(\sum_i P_i M_i \xi_{i0} \right) \right] \\
 &= N^2 \left\{ \sum_i P_i \frac{M_i (N_i - M_i)}{N_i - 1} \sigma_i^2 + \sum_i \left(M_i \xi_{i0} \right)^2 - \left(\sum_i P_i M_i \xi_{i0} \right)^2 \right\} \\
 &= N^2 \left\{ \sum_i P_i \frac{M_i (N_i - M_i)}{N_i - 1} \sigma_i^2 + \sum_i \sum_j P_i P_j M_i M_j \xi_{i0} \xi_{j0} \right. \\
 &\quad \left. - \left(\sum_i P_i M_i \xi_{i0} \right)^2 \right\} = N^2 \left\{ \sum_i P_i \frac{M_i (N_i - M_i)}{N_i - 1} \sigma_i^2 + \sum_i P_i M_i \xi_{i0}^2 \right. \\
 &\quad \left. + \sum_{i \neq j} \sum P_i P_j M_i M_j \xi_{i0} \xi_{j0} - \sum_i P_i M_i \xi_{i0}^2 \right. \\
 &\quad \left. - \sum_{i \neq j} \sum P_i P_j M_i M_j \xi_{i0} \xi_{j0} \right\}
 \end{aligned}$$

$$b_1) \quad \hat{E}(\hat{X}) = N^2 \sum_i \frac{P_i E(M_i)}{N_i} X_i = N^2 \sum_i \frac{P_i n_i}{N_i} X_i$$

$$= N^2 \sum_i \frac{P_i N_i X_i}{N P_i N_i} = \sum_i X_i = \bar{X}$$

$$\begin{aligned}
 b_2) \quad \hat{E}[V(\hat{X})] &= N^2 \left\{ \sum_i P_i \frac{N_i n_i - n_i^2 - V_m}{N_i - 1} \sigma_i^2 + \sum_i P_i n_i \xi_{i0}^2 \right. \\
 &\quad \left. + V_m \sum_i P_i \xi_{i0}^2 + \sum_{i \neq j} \sum P_i P_j n_i n_j \xi_{i0} \xi_{j0} - V_m \sum_i P_i \xi_{i0}^2 \right\}
 \end{aligned}$$

$$\begin{aligned}
 & - \sum_i P_i^2 n_i^2 \xi_{i0} - \sum_i \sum_{i' \neq i} P_i P_{i'} n_i n_{i'} \xi_{i0} \xi_{i'0} \} \\
 & = N^2 \left\{ \sum_i P_i \frac{N_i^2 \xi_{i0}^2}{N_i - 1} \sigma_i^2 + \sum_i P_i \frac{N_i^2 \xi_{i0}^2}{N^2 P_i^2} \right. \\
 & \quad + \sum_i \sum_{i' \neq i} P_i P_{i'} \frac{N_i N_{i'}}{N^2 P_i P_{i'}} \xi_{i0} \xi_{i'0} - \sum_i \frac{P_i N_i \xi_{i0}^2}{P_i^2 N^2} - \sum_i \sum_{i' \neq i} \frac{P_i P_{i'} N_i N_{i'} \xi_{i0} \xi_{i'0}}{N^2 P_i P_{i'}} \} \\
 & + N^2 V_m \left\{ \sum_i P_i \xi_{i0}^2 - \sum_i P_i \xi_{i0}^2 - \sum_i \frac{P_i \sigma_i^2}{N_i - 1} \right\} \\
 & V(X^*) = N^2 \sum_i V(H_i) \xi_{i0}^2 P_i = V_m N^2 \sum_i P_i \xi_{i0}^2
 \end{aligned}$$

En substituant $E[V(\hat{X})]$ et $V(X)^*$ dans $V(\hat{X})$ et simplifiant, nous avons

$$\begin{aligned}
 & V(\hat{X}) = V_1 + V_2 \quad (18) \\
 & V_1 = \sum_i \frac{N_i (N_i P_i - 1) \sigma_i^2}{P_i (N_i - 1)} + \sum_i X_i^2 \left(\frac{1}{P_i} - 1 \right) + \sum_i \sum_{i' \neq i} X_i X_{i'} \left(\frac{P_{ii'}}{P_i P_{i'}} - 1 \right) \\
 & V_2 = N^2 V_m \left[\sum_i P_i \xi_{i0}^2 - \sum_i P_i \frac{\sigma_i^2}{N_i - 1} \right]
 \end{aligned}$$

En pratique le calcul de P_i et de P_{ij} pour les cas où $n > 2$, est très difficile. Nous ferons donc $n = 2$.

7- Plan VI - Stratification suivie de deux degrés de sondage et tirage d'une PSU de chaque strate, avec probabilités inégales de tirage.

La population possède N strates. Dans chaque strate on soumet à l'enquête une seule PSU, avec probabilités inégales p_{ki} pour la

PSU n^0_i de la strate n^0_k (avec $\sum_c p_{ki} = 1$, séparément, pour chaque strate).

Dans chaque PSU ainsi choisie, on tire n_{ki} éléments, de façon qu'en moyenne, 1 sur r des éléments des PSU choisies - soient tirés. Il faut donc que

$$n_{ki} = N_{ki} / \bar{N}_0 r p_{ki} \quad (19)$$

où

$$\bar{N}_0 = \sum_K N_K / N$$

car dans ce cas, nous avons

$$E(n_{ki}) = \frac{1}{N} \sum_K \sum_c \tau_{ki} \frac{N_{ki}}{\bar{N}_0 r p_{ki}} = \frac{N_{00}}{N \bar{N}_0 r}$$

$$= \frac{N_{00}}{r \bar{N}_0} = \frac{\bar{N}_{00}}{r}$$

d'où

$$r = \frac{\bar{N}_{00}}{E(n_{ki})}$$

Un estimateur sans biais du total de la population est

$$\hat{X} = \bar{N}_0 r \sum_K \sum_J x_{ki}$$

Les 4 étapes de calcul sont:

a)

$$E(\hat{X}) = E_1 E_2 (\hat{X}) = \bar{N}_0 r \sum_K E_1 E_2 \sum_J x_{ki}$$

$$= \bar{N}_0 r \sum_K \sum_c \tau_{ki} \frac{p_{ki}}{N_{ki}} \sum_J x_{ki} = r \bar{N}_0 \sum_K \sum_c \tau_{ki} \frac{p_{ki}}{N_{ki}} X_{ki}$$

$$= X^*$$

a2)

$$\begin{aligned}
 V(\hat{X}) &= E(\hat{X} - X^*)^2 = \bar{N}_0^2 E_1 E_2 \left\{ \sum_K \left[\mu_{Kl_0} \bar{x}_{Kl_0} - \mu_{Kl_0} \bar{s}_{Kl_0} + \mu_{Kl_0} \bar{s}_{Kl_0} - \sum_l \mu_{Kl} \mu_{Kl_0} \bar{s}_{Kl_0} \right] \right\}^2 \\
 &= \bar{N}_0^2 E_1 E_2 \left\{ \sum_K \left[\mu_{Kl_0} (\bar{x}_{Kl_0} - \bar{s}_{Kl_0}) + \left[\mu_{Kl_0} \bar{s}_{Kl_0} - E(\mu_{Kl_0} \bar{s}_{Kl_0}) \right] \right] \right\}^2 \\
 &= \bar{N}_0^2 \sum_K \sum_l \sum_{l_0} \mu_{Kl_0}^2 (\bar{x}_{Kl_0} - \bar{s}_{Kl_0})^2 + \bar{N}_0^2 \sum_K E_1 \left[\mu_{Kl_0} \bar{s}_{Kl_0} - E(\mu_{Kl_0} \bar{s}_{Kl_0}) \right]^2 \\
 &\quad + 2 \bar{N}_0^2 \sum_K E_1 \left[\mu_{Kl_0} \bar{s}_{Kl_0} - E(\mu_{Kl_0} \bar{s}_{Kl_0}) \right] \left\{ \sum_l \mu_{Kl} (\bar{x}_{Kl_0} - \bar{s}_{Kl_0}) \right\} \\
 &\quad + \bar{N}_0^2 \sum_{K \neq K'} \left\{ E_1 E_2 \left[\mu_{Kl_0} (\bar{x}_{Kl_0} - \bar{s}_{Kl_0}) \right] \right. \\
 &\quad \left. + E_1 \left[\mu_{Kl_0} \bar{s}_{Kl_0} - E(\mu_{Kl_0} \bar{s}_{Kl_0}) \right] \right\} \left\{ E_1 E_2 \left[\mu_{K'l_0} (\bar{x}_{K'l_0} - \bar{s}_{K'l_0}) \right] \right. \\
 &\quad \left. + E_1 \left[\mu_{K'l_0} \bar{s}_{K'l_0} - E(\mu_{K'l_0} \bar{s}_{K'l_0}) \right] \right\}
 \end{aligned}$$

D'après les raisonnements déjà donnés pour les plans précédents, les doubles-produits sont nuls, et nous avons

$$\begin{aligned}
 V(\hat{X}) &= \bar{N}_0^2 \sum_K \sum_l \mu_{Kl} \frac{\mu_{Kl} (N_{Kl} - \mu_{Kl}) \bar{s}_{Kl}^2}{(N_{Kl} - 1) \mu_{Kl}} \\
 &\quad + \frac{\bar{N}_0^2}{N} \sum_K \sum_l \mu_{Kl} (\mu_{Kl} \bar{s}_{Kl_0} - \sum_l \mu_{Kl} \mu_{Kl_0} \bar{s}_{Kl_0})^2 \\
 &= \bar{N}_0^2 \sum_K \sum_l \mu_{Kl} \left\{ \frac{\mu_{Kl} (N_{Kl} - \mu_{Kl}) \bar{s}_{Kl}^2}{N_{Kl} - 1} \right. \\
 &\quad \left. + \mu_{Kl} \bar{s}_{Kl_0}^2 + \sum_l \mu_{Kl} \mu_{Kl_0} \bar{s}_{Kl_0}^2 - 2 \mu_{Kl} \bar{s}_{Kl_0} \sum_l \mu_{Kl_0} \right\}
 \end{aligned}$$

$$+ \sum_{i \neq c} \sum_{k \neq c} \tau_{ki} \tau_{ki} \tau_{ki} \tau_{ki} \tilde{S}_{kio} \tilde{S}_{kio} \\ - 2 \sum_{i \neq c} \tau_{ki} \tilde{S}_{kio} \tau_{ki} \tau_{ki} \tilde{S}_{kio} \}$$

b₁)

$$E(\hat{X}) = E(X^*) = \bar{N}_0 \sum_K \sum_c \tau_{ki} E(\tau_{ki}) \tilde{S}_{kio}$$

$$= \bar{N}_0 \sum_K \sum_c \tau_{ki} \tau_{ki} \tilde{S}_{kio}$$

$$= \bar{N}_0 \sum_K \sum_c \tau_{ki} \frac{N_{ki}}{\tau_{ki} \bar{N}_0} \tilde{S}_{kio} = \sum_K \sum_c N_{ki} \tilde{S}_{ki}$$

$$= \sum_K \sum_c X_{ki} = \sum_K X_K = X$$

b₂)

$$E[V(\hat{X})] = \bar{N}_0 \sum_K \sum_c \tau_{ki} \left\{ \frac{N_{ki} \tau_{ki} - \tau_{ki} - V_m}{N_{ki} - 1} \right\} \tilde{S}_{kio}^2$$

$$+ (\tau_{ki} + V_m) \tilde{S}_{kio}^2 + \sum_c \tau_{ki}^2 (\tau_{ki} + V_m) \tilde{S}_{kio}^2$$

$$+ \sum_{i \neq c} \sum_{k \neq c} \tau_{ki} \tau_{ki} \tau_{ki} \tau_{ki} \tilde{S}_{kio} \tilde{S}_{kio} - 2 (\tau_{ki} + V_m) \tilde{S}_{kio}^2 \tau_{ki}$$

$$- 2 \sum_{i \neq c} \tau_{ki} \tilde{S}_{kio} \tau_{ki} \tau_{ki} \tilde{S}_{kio} \}$$

$$= \bar{N}_0 \sum_K \sum_c \tau_{ki} \left\{ \frac{N_{ki} \tau_{ki} - \tau_{ki}}{N_{ki} - 1} \right\} \tilde{S}_{kio}^2$$

$$+ (\tau_{ki} \tilde{S}_{kio} - \sum_c \tau_{ki} \tau_{ki} \tilde{S}_{kio}) \}$$

$$\begin{aligned}
 & + \bar{N}_0^2 V_m \sum_K \sum_L \tau_{KL} \left\{ -\frac{\sigma_{KL}^2}{N_{KL}-1} + \sum_{i,j} \tau_{Ki} \bar{S}_{Kio} + \bar{S}_{Kio}^2 - 2 \sum_{i,j} \tau_{Kio} \tau_{Kij} \right\} \\
 & = \bar{N}_0^2 \sum_K \sum_L \tau_{KL} \left\{ \frac{\bar{N}_0 \sum_i \tau_{Ki} - \frac{N_{KL}^2}{N_{KL}}}{N_{KL}-1} \sigma_{KL}^2 + \left(\frac{N_{KL} \bar{S}_{Kio}}{\bar{N}_0 \sum_i \tau_{Ki}} - \sum_{i,j} \tau_{Kij} \frac{N_{KL} \bar{S}_{Kio}}{\bar{N}_0 \sum_i \tau_{Ki}} \right) \right\} \\
 & + \bar{N}_0^2 V_m \sum_K \sum_L \tau_{KL} \left(-\frac{\sigma_{KL}^2}{N_{KL}-1} + \sum_{i,j} \tau_{Ki} \bar{S}_{Kio} + \bar{S}_{Kio}^2 - 2 \sum_{i,j} \tau_{Kio} \tau_{Kij} \right) \\
 & = \sum_K \sum_L \frac{(\bar{N}_0 \sum_i \tau_{Ki} - 1) N_{KL} \sigma_{KL}^2}{\tau_{KL} (N_{KL} - 1)} + \\
 & \quad + \sum_K \sum_L \tau_{KL} \left(\frac{X_{KL}}{\tau_{KL}} - X_K \right)^2 \\
 & + V_m \bar{N}_0^2 \sum_K \sum_L \tau_{KL} \left(-\frac{\sigma_{KL}^2}{N_{KL}-1} + \sum_{i,j} \tau_{Ki} \bar{S}_{Kio} + \bar{S}_{Kio}^2 - 2 \sum_{i,j} \tau_{Kio} \tau_{Kij} \right) \\
 & V(X^*) = \sum_K \sum_L \bar{N}_0^2 \tau_{KL} V(\mu_{KL}) \\
 & = V_m \bar{N}_0^2 \sum_K \sum_L \tau_{KL} \left[\sum_{i,j} \tau_{Ki} \bar{S}_{Kio} - \sum_{i,j} \tau_{Kij} \frac{\sigma_{KL}}{N_{KL}-1} \right]
 \end{aligned}$$

En substituant les valeurs de $\sum V(\hat{X})$ et $V(X^*)$ dans $V(\hat{X})$, nous avons

$$\begin{aligned}
 & V(\hat{X}) = V_1 + V_2 \quad (20) \\
 & V_1 = \sum_K \sum_L \frac{(\bar{N}_0 \sum_i \tau_{Ki} - 1) N_{KL} \sigma_{KL}^2}{\tau_{KL} (N_{KL} - 1)} + \sum_K \sum_L \tau_{KL} \left(\frac{X_{KL}}{\tau_{KL}} - X_K \right)^2 \\
 & V_2 = V_m \bar{N}_0^2 \left[\sum_K \sum_L \tau_{KL} \sum_{i,j} \tau_{Ki} \bar{S}_{Kio} - \sum_K \sum_L \tau_{KL} \sum_{i,j} \tau_{Kij} \frac{\sigma_{KL}}{N_{KL}-1} \right]
 \end{aligned}$$

8- Plan VII- Sondage à deux degrés, le premier avec probabilités inégales et tirages bernoulliens.

La population a N PSU dont celle qui porte le numero i contient N_i éléments. On se propose de soumettre à l'enquête n PSU tirées avec probabilités inégales et tirages bernoulliens. Dans chaque PSU ainsi choisie on soumet à l'enquête n_i éléments obtenus par tirages exhaustifs et avec probabilités égales.

Le tirage des n_i éléments dans la PSU n°i est supposé indépendant du tirage des PSU et aussi du tirage des éléments dans les autres PSU.

Posons la probabilité de choix de la PSU n°i égale à p_i , et l'effectif d'échantillon dans cette PSU égal à

$$n_i = N_i/p_i r \quad (21)$$

Un estimateur sans biais de X, total relatif à la population, est

$$\hat{X} = \frac{r}{n} \sum_i \sum_j x_{ij} \quad (22)$$

a1)

$$\begin{aligned} E(\hat{X}) &= E\left(\frac{r}{n} \sum_i \sum_j x_{ij}\right) = \frac{r}{n} \sum_i E_1 \sum_j E_2 x_{ij} = \frac{r}{n} \sum_i \frac{N_i}{N_i} \sum_j x_{ij} \\ &= \frac{r}{N} \sum_i N_i \frac{N_i}{N_i} X_i = \frac{r}{n} \sum_i N_i \frac{N_i}{N_i} \bar{x}_{i0} = X^* \end{aligned}$$

a2)

$$\begin{aligned} V(\hat{X}) &= E\left(\frac{r}{n} \sum_i \sum_j x_{ij} - \frac{r}{n} \sum_i N_i \bar{x}_{i0} + \frac{r}{n} \sum_i N_i \bar{x}_{i0} - \frac{r}{n} \sum_i N_i \bar{x}_{i0}\right)^2 \\ &= \left(\frac{r}{n}\right)^2 E\left\{\sum_i N_i (\bar{x}_{i0} - \bar{x}_{i0}) + N[\bar{N} \bar{x}_{i0} - E(N_i \bar{x}_{i0})]\right\}^2 \end{aligned}$$

$$= \left(\frac{2}{n}\right) \left\{ \sum_i \mu_i^2 \mathcal{E}_2(\bar{x}_{i0} - \xi_{i0})^2 + \sum_{i \neq i'} \mu_i \mu_{i'} \mathcal{E}_2(\bar{x}_{i0} - \xi_{i0})(\bar{x}_{i'0} - \xi_{i'0}) \right. \\ \left. + \mathcal{E}_1[\mu_i \bar{\xi}_{i0} - \mathcal{E}(\mu_i \xi_{i0})] + 2 \mu_i \mathcal{E}_1[\mu_{i'} \bar{\xi}_{i'0} - \mathcal{E}(\mu_{i'} \xi_{i'0})][\mu_i \mathcal{E}_2(\bar{x}_{i0} - \xi_{i0})] \right\}$$

Bien que les tirages du premier degré de sondage soient bernoulliens, comme les tirages du deuxième degré sont indépendants les uns des autres, les doubles produits sont nuls. Nous avons

$$V(\hat{X}) = \left(\frac{2}{n}\right) \left\{ \sum_i \mu_i \mathcal{I}_i \frac{\mu_i (N_i - \mu_i) \sigma_i^2}{N_i - 1} + \mu \sigma_{\mu_i}^2 \xi_{i0} \right\} \\ = \left(\frac{2}{n}\right) \left\{ \sum_i \mu_i \mathcal{I}_i \frac{\mu_i (N_i - \mu_i) \sigma_i^2}{N_i - 1} + \mu [\mathcal{E}(\mu_i^2 \xi_{i0}^2) - (\mathcal{E} \mu_i \xi_{i0})^2] \right\} \\ = \left(\frac{2}{n}\right) \left\{ \sum_i \mu_i \mathcal{I}_i \frac{\mu_i (N_i - \mu_i) \sigma_i^2}{N_i - 1} + \mu \sum_i \mathcal{I}_i \mu_i^2 \xi_{i0}^2 - \mu \left(\sum_i \mathcal{I}_i \mu_i \xi_{i0} \right)^2 \right\}$$

b1)

$$E(\hat{X}) = \mathcal{E}(X^*) = \frac{2}{n} E(\mu) \sum_i \mathcal{I}_i \frac{E(\mu_i)}{N_i} X_i = \frac{2}{n} \sum_i \mathcal{I}_i \frac{n_i}{N_i} X_i \\ = 2 \sum_i \mathcal{I}_i \frac{N_i}{\mathcal{I}_i \cdot 2} X_i = \sum_i X_i = X$$

b2)

$$\mathcal{E} V(\hat{X}) = \frac{2}{N^2} \left\{ \sum_i E(\mu) \mathcal{I}_i \frac{N_i E(\mu_i) - E(\mu_i^2)}{N_i - 1} \sigma_i^2 \right. \\ \left. + E(\mu) \sum_i \mathcal{I}_i E(\mu_i^2) \xi_{i0}^2 - E(\mu) \left[\sum_i \mathcal{I}_i E(\mu_i^2) \xi_{i0}^2 \right] \right. \\ \left. - E(\mu) \sum_{i \neq i'} \mathcal{I}_i \mathcal{I}_{i'} E(\mu_i) E(\mu_{i'}) \xi_{i0} \xi_{i'0} \right\}$$

$$\begin{aligned}
 &= \left(\frac{\varepsilon}{n}\right)^2 \left\{ \sum_i \tau_i n_i \frac{N_i n_i - n_i^2 - V_m}{N_i - 1} \sigma_i^2 + n \sum_i \tau_i (n_i + V_m) \xi_{i0}^2 \right. \\
 &\quad \left. - n \sum_i \tau_i (n_i + V_m) \xi_{i0}^2 - n \sum_{i \neq i'} \tau_i \tau_{i'} n_i n_{i'} \xi_{i0} \xi_{i'0} \right\} \\
 &= - \frac{V_m^2}{n} \sum_i \tau_i \frac{\sigma_i^2}{N_i - 1} + \frac{\varepsilon}{n} V_m \sum_i \tau_i \xi_{i0}^2 - \frac{\varepsilon}{n} V_m \sum_i \tau_i \xi_{i0}^2 \\
 &\quad + \frac{\varepsilon}{n} \sum_i \frac{N_i}{\varepsilon \tau_i} \left(\frac{N_i}{\varepsilon \tau_i} \right)^2 \sigma_i^2 + \frac{\varepsilon}{n} \sum_{i \neq i'} \left[\frac{N_i \xi_{i0}}{\varepsilon \tau_i} - \sum_{i'} \tau_{i'} \frac{N_{i'} \xi_{i'0}}{\varepsilon \tau_{i'}} \right]^2
 \end{aligned}$$

Nous avons trouvé cette dernière expression en remplaçant n_i par N_i/rp_i et séparant les termes contenant V_m et les autres.

$$V(X^*) = E \left[\frac{\varepsilon}{n} \mu \sum_i \tau_i \cdot \mu_i \cdot \xi_{i0}^2 - \frac{\varepsilon}{n} \mu \sum_i \tau_i n_i \xi_{i0}^2 + \frac{\varepsilon}{n} \mu \sum_i \tau_i n_i \xi_{i0}^2 - \frac{\varepsilon}{n} n \sum_i \tau_i n_i \xi_{i0}^2 \right]^2$$

ou, comme dans les autres cas, nous avons ajouté et retranché un même terme; quant au dernier terme il est égal à X . En considérant, qu'ici aussi les doubles produits sont nuls pour les mêmes raisons que nous avons données plus haut, nous avons

$$\begin{aligned}
 V(X^*) &= \frac{\varepsilon}{n^2} E \left[\mu \sum_i \tau_i \xi_{i0}^2 (1 - n_i) + (1 - n) \sum_i \tau_i n_i \xi_{i0}^2 \right]^2 \\
 &= \frac{\varepsilon}{n^2} \left\{ E(\mu^2) \sum_i \tau_i \xi_{i0}^2 V(\mu) + V(\mu) \left(\sum_i \tau_i n_i \xi_{i0}^2 \right)^2 \right\} \\
 &= \frac{\varepsilon}{n^2} \left\{ (n + V_m) \sum_i \tau_i \xi_{i0}^2 V_m + V_m \left(\sum_i \tau_i n_i \xi_{i0}^2 \right)^2 \right\} \\
 &= \varepsilon V_m \sum_i \tau_i \xi_{i0}^2 + \frac{\varepsilon V_m}{n^2} \sum_i \tau_i \xi_{i0}^2 + \frac{V_m \varepsilon}{n^2} \left(\sum_i \tau_i \frac{N_i}{\varepsilon \tau_i} \xi_{i0}^2 \right)^2
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{e^2}{n} V_m \sum_i \frac{x_i^2}{N_i - 1} + \frac{4V_m}{n^2} \sum_i x_i^2 \bar{x}_i + \frac{V_m}{n^2} \left(\sum_i N_i \bar{x}_i \right)^2 \\
 &= \frac{e^2}{n} V_m \sum_i \frac{x_i^2}{N_i - 1} + \frac{4V_m}{n^2} \sum_i x_i^2 \bar{x}_i + \frac{V_m}{n^2} X^2
 \end{aligned}$$

En substituant $\bar{x}^*(X)$ et $V(X^*)$ dans $V(\hat{X})$, nous obtenons comme des termes contenant V_m (somme que nous désignerons par V_2):

$$\begin{aligned}
 V_2 = & - \frac{e^2}{n} V_m \sum_i \frac{x_i^2}{N_i - 1} + \frac{V_m}{n} \sum_i x_i^2 \bar{x}_i \\
 & - \frac{V_m}{n} \sum_i \frac{x_i^2}{N_i - 1} + \frac{e^2}{n} V_m \sum_i \frac{x_i^2}{N_i - 1} \\
 & + \frac{V_m}{n^2} \sum_i x_i^2 \bar{x}_i + \frac{V_m}{n^2} X^2
 \end{aligned}$$

Nous ne pouvons pas la simplifier. Mais les termes 1, 3 et 5 sont négligeables à côté des 3 autres termes. Nous allons donc les supprimer, et nous aurons

$$\begin{aligned}
 \left[\begin{aligned}
 \hat{V}(X) &= V_1 + V_2 \\
 V_1 &= \frac{1}{n} \sum_i \frac{N_i (x_i - 1) \bar{x}_i^2}{N_i (N_i - 1)} + \frac{1}{n} \sum_i x_i^2 \left(\frac{x_i}{N_i} - X \right)^2 \\
 V_2 &= \frac{V_m X^2}{n^2} + \frac{V_m}{n} \sum_i x_i^2 \bar{x}_i + V_m \sum_i x_i^2 \bar{x}_i
 \end{aligned} \right. \quad (23)
 \end{aligned}$$

Telles sont les nouvelles formules d'estimations et de variances pour nos divers plans de sondage, formules qui vont être utilisées à présent dans les 3^o et 4^o parties.

9- La comparaison des différents plans de sondage au point de vue de l'augmentation de variance

Les formules des V_2 , que nous avons trouvées au cours de la deuxième partie, sont assez compliquées pour être comparables

sous leurs formes présentes. Afin de les comparer, nous calculons les V_2 des différents plans pour un cas tout à fait abstrait. Nous changeons les paramètres de la population et les rendons tels que les V_1 de tous les plans soient égales entre elles. Le procédé le plus simple est que nous rendions toutes les V_1 égales à zéro. Cela veut dire que nous appliquerons les formules de V_2 à une population où (1) les effectifs de toutes les unités de même ordre sont égaux, et (2) les mesures de tous les éléments sont aussi égales.

En fin, afin de simplifier les formules, nous posons:

- 1) la fraction de sondage égale à $1/R$.
- 2) La fraction de sondage pour chaque degré (dans un sondage à deux degrés) égale à $1/\sqrt{R}$;
- 3) L'effectif total de la population égal à M ;
- 4) dans le cas d'un double découpage de la population, qui peut être fait en vue d'un sondage stratifié ou d'un sondage à deux degrés:

Le nombre des classes d'ordre un, égal à $\sqrt{M}$;

Le nombre des éléments dans chaque classe égal à $\sqrt{M}$;

- 5) dans le cas d'un triple découpage (stratifications et degrés de sondage):

Le nombre des classes de 1^o ordre égal à $\sqrt[3]{M}$;

Le nombre des classes de deuxième ordre, dans chaque classe de premier ordre, égal à $\sqrt[3]{M}$;

Le nombre des éléments, dans chaque classe de deuxième ordre, égal à $\sqrt[3]{M}$;

Nous éliminons les termes négligeables des V_2 .

Dans le cas des sondages avec probabilités inégales, nous étudions les 3 cas suivants, où intervient une lettre r qui a le sens suivant:

$$\text{Echantillon de } \left[\begin{array}{l} 2 \text{ PSU Avec tirages Exhaustifs, } R = r N/2 = r \sqrt{M}/2 \\ \text{une PSU de chaque strate, } R = N \cdot r/N = r \sqrt[3]{M} \\ n \text{ PSU avec tirages bernoulliens, } R = r/N \end{array} \right.$$

Afin de compléter nos comparaisons, nous calculons également V_2 pour un plan de sondage simple (une seule strate), en négligeant le fait que, dans ce cas, la théorie valable pour des nombres de strates grands ne s'applique pas (voir §§ 2, 3 et 4 de la première partie).

En utilisant les approximations et remarques ci-dessus, nous pouvons écrire les V_2 sous la forme (générale) suivante:

$$V_2 = Y V_m X^2 \quad (24)$$

où $V_m = V(\mu)$, $X =$ le total de la population et où les Y sont donnés dans le tableau suivant:

Plan de sondage	Nombre de découpages	Valeur de Y	Expression de Y en fonction de V
Simple	1	$(R/M)^2$	V
Plan I-Stratification simple avec probabilités égales	2	$R^2/M \sqrt{M}$	$V \sqrt{M}$
Plan II-Sondage à deux degrés avec probabilités égales	2	$R \sqrt{R/M} \sqrt{M}$	$V \sqrt{M/R}$

Plan de sondage	Nombre de découpages	Valeur de Y	Expression de Y en fonction de V
Plan III-Stratification suivie de deux degrés de sondage avec probabilités égales	3	$R\sqrt{R/M}\sqrt[3]{M} + R/M$	$V\frac{\sqrt[3]{2}}{\sqrt{R}} + \frac{M}{R}$
Plan IV-Stratification des PSU (considérée comme découpage de la population)	3	$\frac{R\sqrt{R}}{M\sqrt[3]{M}}$	$V\sqrt[3]{M^2}/\sqrt{R}$
Plan V-2 PSU avec probabilités inégales et tirages exhaustifs	2	$8(R/M)^2$	8V
Plan VI-Stratification, une PSU de chaque strate, probabilités inégales	3	$(R/M)^2\sqrt[3]{M^2}$	$V\sqrt[3]{M^2}$
Plan VII-Tirage bernoullien des PSU avec probabilités inégales	2	$\frac{R}{M}(1 + \sqrt{\frac{R}{M}} + \frac{1}{\sqrt{M}})$	$V(\frac{M}{R} + \sqrt{\frac{M}{R}} + \frac{\sqrt{M}}{R})$

Dans le tableau ci-dessus, nous avons exprimé les Y de chaque plan en fonction de celui V du plan simple.

En remarquant que $M > R$, nous tirons des données du tableau les deux résultats suivants:

a) Plus le nombre des découpages de la population est grand (pour stratification ou pour sondage) et plus V_2 sera grande.

b) Plus le nombre d'unités échantillon de chaque ordre est petit, et plus V_2 sera grande.

Donc, pour diminuer V_2 nous devons (1) découper la population le moins de fois possible, et (2) rendre le nombre d'unités échantillon aussi grand que possible.

La condition (1) est défavorable pour V_1 si le but du découpage est une stratification, mais favorable pour V_1 si le but du découpage est de créer un degré de sondage de plus. La condition (2) est toujours favorable pour V_1 (aussi bien que pour V_2).

On pourrait très facilement comparer les plans de sondage d'après le tableau ci-dessus.

TROISIEME PARTIE

REPARTITION OPTIMUM D'ECHANTILLON QUAND LA POPULATION A UNE STRUCTURE DONNEE.

1- Enoncé général du problème.-

Nous avons étudié, au cours de la seconde partie, l'augmentation de la variance sous l'effet des erreurs d'arrondissement.

Nous allons à présent résoudre un certain nombre de problèmes de répartition-optimum de l'échantillon quand la population a une structure donnée; et nous laissons l'étude des problèmes concernant la recherche des structures optimum de la population pour la quatrième partie.

o
o o

Voici ce qu'on entend par "répartition optimum".

On a une population avec une structure donnée où l'on veut exécuter un plan donné de sondage. L'estimateur est $\hat{X}(n)$ où $n = [n, \dots, n_x, \dots]$ a le même sens que dans le §1 de la deuxième partie. Les fonctions de variance et de coût sont, respectivement, $V(\hat{X})$ et $c(n)$. On se propose de calculer les n tels que (1) ou bien V soit minimum pour une valeur fixe de c , (2) ou bien c soit minimum pour une valeur fixe V_0 de V . Tels sont les deux points de vue symétriques de la théorie classique.

Mais il est essentiel de distinguer d'abord deux types de problèmes très différents;

Type 1: Abstraction faite de toute erreur d'arrondissement, certaine théorie classique procure une répartition optimum de l'échantillon, c'est-à-dire fournit, pour les (n) , des valeurs théoriques qui généralement ne sont pas entières. Or, il est absolument nécessaire d'arrondir ces valeurs (n) pour pouvoir définir l'échantillon à tirer; ceci introduit un accroissement de variance qu'on va apprendre à calculer et à utiliser pour amender la théorie classique.

Type 2 : Pour des raisons de commodité (comme on l'a déjà vu) on peut avoir été conduit à rechercher, au lieu d'un effectif optimum d'échantillon n_x séparément pour chaque unité, une fraction unique de sondage optimum pour toutes les unités du même ordre, ce qui nous éloigne du point optimum proprement dit, mais, en revanche, simplifie beaucoup les calculs mathématiques.

L'existence simultanée de ces deux types de problèmes pouvait paraître à première vue surprenante. En réalité les problèmes traités ne sont pas exactement les mêmes. Dans les problèmes du type 2 on ne recherche pas l'échantillon optimum proprement dit, mais celui qui est optimum parmi une catégorie d'échantillons très limitée.

o
o o

Liste des problèmes traités ici:

Problèmes du type 1-

1- Point de vue c constant.

1.1 Le problème " de Neyman ". Pour un simple plan de sondage stratifié, et une fonction de coût

$$C = b \sum n_i c_i \quad (a = 0 \text{ voir } \S 2)$$

on se place au premier point de vue. Neyman a trouvé une "repartition optimum" classique (1934), théorème que Yates a étendu (1935) au cas de

$$C = \sum b_i n_i c_i$$

On se limite au calcul de l'accroissement de variance, consécutif à la transformation des n théoriques en μ entiers.

1.2 Le " problème de Neyman " pour un plan de sondage stratifié et tirage bernoullien, avec probabilités inégaes dans chaque strate.

2° -Point de vue: V constant : L'équivalent du problème de Neyman, avec un échantillon, non plus stratifié, mais à deux degrés.

- Extension de ce problème pour k degrés de sondage.
- Même problème avec tirage exhaustif (2 PSU seulement) et probabilités inégaes au 1° degré.
- Même problème avec stratification
 - deux degrés de sondage
 - 1 PSU de chaque strate
 - Probabilités inégales.

- Même problème avec

2 degrés de sondage
n PSU (tirage bernoullien)
Probabilités inégales.

Problèmes du type 2.

Contrairement aux problèmes du type 1, nous ne faisons pas ici abstraction des erreurs d'arrondissement quand nous calculons les valeurs optimum des effectifs de sondage, mais nous utilisons les formules de variance déjà calculées dans la deuxième partie pour les différents plans de sondage, en tenant compte des erreurs d'arrondissement.

Les problèmes traités correspondent aux plans II, III, IV, V, VI, et VII. Il faut distinguer deux cas.

1-Plan II, III, IV, et VII.

Dans une étape préliminaire où on ne tient pas compte des erreurs d'arrondissement, on se donne comme objectif une certaine variance V_0 , et on détermine le coût minimum c_0 correspondant (par exemple V_0 devra être 5% ou 10% du total); soit r_0, s_0 les valeurs optimum de $r, s, \dots$ correspondantes.

Après quoi, pour des raisons de commodité de calcul on abandonne le point de vue V constante pour le point de vue c constant; c'est-à-dire que la suite des calculs est faite en supposant le coût constant et égale à c_0 , et en cherchant les valeurs de $r, s, \dots$ qui rendent minimum la variance corrigée des erreurs d'arrondissement des effectifs.

Ce changement de point de vue est en fait sans importance; et (sauf anomalies peu vraisemblable) tout se passe comme si

l'on avait déterminé au début la variance la plus petite V_0 correspondant au coût donné c_0 (abstraction faite des phénomènes d'arrondissement). L'intérêt de cette façon de commencer les calculs est qu'elle fournit de la variance V une limite inférieure V_0 qui est aussi un ordre de grandeur de V . Pour un pays comme L'Iran la précision V_0 a un sens plus concret que le coût c_0 . Car on n'a pas en général à gérer un budget, mais à proposer telle enquête qui (si elle est acceptée) devra être effectuée avec le minimum d'argent.

2. Plan V et VI.

Il s'agit de deux cas où l'on a à choisir une seule fraction de sondage.

On ne peut alors trouver "d'optimum", ni pour une valeur fixe de V , ni une valeur fixe de c . On peut calculer seulement une valeur optimum indépendante de notre décision à propos de V et de c . Nous utiliserons le terme "optimum indépendant" dans les § 5.4 et 5.5 pour désigner la fraction de sondage qui rend le produit cV minimum.

Marché générale des calculs pour les problèmes de type 1.

a) Nous calculons les valeurs optimum de n sans tenir compte des erreurs d'arrondissement. En même temps nous donnons les formules classiques de $\hat{X}(n)$, $V[\hat{X}(n)]$, et V_{opt} , et aussi la fonction de coût (en tenant compte de ce que nous allons en dire au § 2). V_{opt} ci-dessus mentionné, désigne la valeur $V[\hat{X}(n)]$ quand on y substitue les valeurs optimum calculées des n .

a2) Nous arrondissons (d'après une des méthodes d'arrondissement) les n optimum aux valeurs $= [n]$ ou $[n] + 1$. Mathématiquement, nous fixons la valeur de $V_m = V(\mu)$; car au point de vue de nos calculs, c'est la seule différence qui existe entre les méthodes.

a3) Nous calculons $V[\hat{X}(\mu)] = E\{V[\hat{X}(\mu)]\}$ ce qui donne l'équation:

$$V[\hat{X}(\mu)] = Vt \quad (1)$$

où $V = V_{opt}$, et t est une fonction de V et des autres paramètres de la population; et nous allons vérifier au cours de nos calculs que t est une fonction non-décroissante de V et est toujours plus grande que un.

a4) Après avoir calculé l'expression de t , un premier résultat est immédiat: l'accroissement relatif de la variance, cause par les erreurs d'arrondissement, est

$$d = t - 1 \quad (2)$$

a5) Nous devons calculer la valeur W de V , telle que V soit égale à V_0 ; nous substituons dans la formule (1) $V[\hat{X}(\mu)]$ par la valeur fixe V_0 et V par W ; il vient

$$W t(W) = V_0 \quad (3)$$

équation en W .

Pratiquement ici nous avons deux cas: (1) si l'équation (3) en W est algébriquement soluble, nous n'avons plus de problème. Mais si elle ne l'est pas, nous devons la résoudre par itération.

Grâce aux propriétés de t et à la proposition que nous allons démontrer ci-dessous, nous donnons une méthode de résolution de l'équation (1) par itération.

Proposition - Soit V_i , t_i et V_i' des variables reliées entre elles par les formules récurrentes:

$$V_i' = t_i V_i \quad (4)$$

$$t_i = \mathcal{P}(V_i) \quad (5)$$

$$V_i t_i = V_0 / t_i \quad (6)$$

où $i = 1, 2, 3, \dots$; V_0 est une constante.

$$\frac{d\mathcal{P}(V_i)}{dV_i} > 0 \quad (7)$$

$$\mathcal{P}(V_i) > 1 \quad (8)$$

Pour toutes les valeurs impaires de $i = 2n + 1$, on a

$$t_{2n} > t_{2n+1} > t_{2n-1} \quad \text{et} \quad V_{2n-1}' < V_{2n+1}' < V_0 \quad (9)$$

et pour toutes les valeurs paires de $i = 2n + 2$:

$$t_{2n} > t_{2n+2} > t_{2n+1} \quad \text{et} \quad V_{2n}' > V_{2n+2}' > V_0$$

C'est-à-dire que les valeurs de V_i' oscillent autour de V_0 , et s'en rapprochent à mesure que i croît.

En effet on peut écrire (6) sous la forme

$$V_{i+1}' = \frac{V_i t_{i-1}}{t_i} \quad (10)$$

et (4) sous la forme

$$V_i' = \frac{V_0 t_i}{t_i - 1} \quad (11)$$

D'abord nous démontrons la proposition pour $i = 0, 1, 2$ et 3, à l'aide d'une suite de propositions que nous portons dans le tableau suivant:

n°	Propositions	Démonstrations
(a)	$t_0 > 1$	d'après (8)
(b)	$V_0' > V_0$	d'après (a) et (4), $V_0' = t_0 V_0 > V_0$
(c)	$V_1 < V_0$	d'après (6) et (a), $V_1 = \frac{V_0}{t_0} < V_0$
(d)	$t_1 < t_0$	d'après (7) et (c)
(e)	$V_1' < V_0$	d'après (11), $V_1' = \frac{V_1 t_0}{t_1} < V_0$ d'après (d)
(f)	$V_1 < V_2 < V_0$	d'après (10), $V_2 = \frac{V_1 t_0}{t_1} > V_1$ d'après (d) et d'après (6) $V_2 = \frac{V_0}{t_1} < V_0$ d'après (8)
(g)	$t_1 < t_2 < t_0$	d'après (7) et (f)
(h)	$V_0 < V_2 < V_0'$	d'après (11), $V_2' = \frac{V_2 t_2}{t_1} > V_2$ d'après (9), et comme d'après (4), $V_0' = t_0 V_0$ $\frac{V_2}{V_0 t_2} = \frac{V_0 t_2}{t_1 t_2 V_0}$ $= \left(\frac{1}{t_0} \right) \left(\frac{t_1}{t_2} \right) < 1$ d'après (3)

n°	Propositions	Démonstrations
(i)	$V_1 < V_3 < V_2$	d'après (10) $V_3 = \frac{V_2 t_1}{t_2} < V_2$ d'après (9); et d'après (6) $V_3 = \frac{V_0}{t_2} = \frac{V_1 t_0}{t_2} > V_1$
(j)	$t_1 < t_3 < t_2$	d'après (7) et (i),
(k)	$V_1' < V_3' < V_0$	d'après (11) $V_3' = \frac{V_2 t_1}{t_2} < V_0$ d'après (j), et comme d'après (11) $V_1' = \frac{V_2 t_1}{t_0}$ $\frac{V_3'}{V_1'} = \frac{V_0 t_3}{t_2} \frac{t_0}{V_0 t_1}$ $= \left(\frac{t_3}{t_1} \right) \left(\frac{t_0}{t_2} \right) > 1$ d'après (g) et (j)

Les inégalités (a), (d), (e), (g), (h), (j), et (k) démontrent la proposition pour $i = 0, 1, 2, 3$

Alors nous démontrons, à l'aide du tableau suivant, que si les inégalités

$$t_{2n-3} < t_{2n-1} < t_{2n-2} \quad (12)$$

sont vérifiées, les inégalités suivantes le sont également:

$$t_{2n-1} < t_{2n} < t_{2n-2} \quad (13)$$

$$t_{2n-1} < t_{2n+1} < t_{2n} \quad (14)$$

$$V_0 < V_{2n} < V_{2n-2} \quad (15)$$

$$V_0 > V_{2n+1} > V_{2n-1} \quad (16)$$

n°	Proposition	Démonstration
(a')	$V_{2n-1} < V_{2n} < V_{2n-2}$	d'après (6), $\frac{V_{2n}}{V_{2n-1}} = \frac{V_{2n-2}}{t_{2n-1} V_0} \text{ et } \frac{V_{2n}}{V_{2n-2}} = \frac{V_0 t_{2n-3}}{t_{2n-1} V_0} \text{ ce qui}$ donne avec (12) les inégalités (a)
(b')	$t_{2n-1} < t_{2n} < t_{2n-2}$	d'après (7) et (a)
(c')	$V_0 < V_{2n} < V_{2n-2}$	d'après (11) $V_{2n} = \frac{V_0 t_{2n}}{t_{2n-1}} > V_0 \text{ d'après (b)}$ et $\frac{V_{2n}}{V_{2n-2}} = \frac{V_0 t_{2n} t_{2n-3}}{t_{2n-1} V_0 t_{2n-2}} = \left(\frac{t_{2n}}{t_{2n-2}} \right) \left(\frac{t_{2n-3}}{t_{2n-1}} \right) < 1$
(d')		d'après (6) $\frac{V_{2n+1}}{V_{2n-1}} = \frac{V_0 t_{2n-2}}{t_{2n} V_0} \text{ et}$

N°	Proposition	Démonstration
(e') (f')	$t_{2n-1} < t_{2n+1} < t_{2n}$ $V_0 > V'_{2n+1} > V'_{2n-1}$	$\frac{V_{2n+1}}{V_{2n}} = \frac{V_0 t_{2n-1}}{t_{2n} V_0}, \text{ et}$ d'après (b) les inégalités (d) sont immédiates d'après (7) et (d') d'après (11) $V'_{2n+1} = \frac{V_0 t_{2n+1}}{t_{2n}} < V_0$ d'après (é); et $\frac{V'_{2n+1}}{V'_{2n-1}} = \left(\frac{V_0 t_{2n+1}}{t_{2n}} \right) \left(\frac{t_{2n-2}}{V_0 t_{2n-1}} \right)$ $= \left(\frac{t_{2n+1}}{t_{2n-1}} \right) \left(\frac{t_{2n-2}}{t_{2n}} \right) > 1$ d'après (B') et (é)

Les inégalités (b'), (c'), (e'), et (f') sont les mêmes que les inégalités (13), (14), (15) et (16) que nous voulions démontrer.

Ainsi la proposition est bien démontrée par récurrence.

Application à l'équation (1). Dans le cas où l'équation (1) n'a pas une forme algébrique simple, d'abord nous donnons la valeur V_0 à W (non pas à $V[\hat{X}(P)]$) et ainsi nous calculons t_0 et V'_0 . D'après la proposition qu'on vient d'établir

on a $V_0' > V_0$

Nous calculons alors $V_1 = V_0/t_0$, t_1 et V_1' . Cette fois $V_1 < V_0$. Nous calculons donc $V_2 = V_0/t_1$, t_2 et V_2' qui est plus grande que V_0 mais plus petite que V_1' . Et ainsi de suite; nous continuons ce procédé et nous trouvons les valeurs V_0' , V_2' , V_4' qui s'approchent à droite à V_0 , et les valeurs V_1' , V_3' , V_5' qui s'approchent à gauche à V_0 , et, de cette façon, nous pouvons trouver avec l'exactitude voulue la valeur V_1 que nous devons donner à V pour que la variance de l'estimation

$$V[\hat{X}(t)]$$

soit égale à V_0 .

Marche Générale des calculs pour les problèmes du type 2-

1b) Dans les problèmes du deuxième type où nous calculons les effectifs de sondage en tenant compte des erreurs d'arrondissement, nous utiliserons les formules $V = V_1 + V_2$ trouvées pour chaque plan, dans la première partie. Ici nous calculons les valeurs de r et s telles que le coût de sondage soit minimum, pour la valeur fixe de V_0 de la variance.

La raison pour que nous considérons les r et s , et non pas les n_{α} comme dans le cas des problèmes du type 1, est qu'ici nous voulons utiliser les calculs faits au cours de la seconde partie, et puis les continuer au cours de la quatrième partie (où l'utilisation des n_{α} au lieu des r et s , est pratiquement impossible);

tandis que, dans le premier type de problèmes, l'utilisation des n_x est possible; et on sait que dans ce cas, nous approchons - beaucoup plus à la valeur optimum absolue.

Ici se posent deux problèmes:

1b1) quand on tient compte des erreurs d'arrondissement, quel est le nouvel optimum?

1b2) en quoi diffère-t-il de l'ancien (quand nous ne tenons pas compte des erreurs d'arrondissement)?

Pour résoudre ces problèmes, nous effectuerons les étapes suivantes de calcul:

b1) Nous calculons les valeurs r_0 et $s_0, \dots$ de r et $s, \dots$ telles que, pour une valeur fixe de la variance, V_0 , le coût du sondage soit minimum, quand nous ne tenons pas compte des erreurs d'arrondissement. Désignons cette valeur du coût par c_0 . Dans cette étape de calcul, nous utilisons les formules de V_1 calculées dans la première partie, c'est-à-dire nous calculons r et s d'après la formule $V_1 + \lambda c = \text{minimum}$ ou d'autres formules équivalentes.

b2) Nous calculons la valeur $V'_0 = V_{01} + V_{02}$ en substituant r_0 et s_0 dans la formule $V = V_1 + V_2$, ce qui nous donne la valeur de $V \left[\hat{X}(\mu) \right]$ quand nous n'avons pas tenu compte des erreurs d'arrondissement, dans le calcul de r et $s, \dots$

Cette étape de calcul correspond aux problèmes du type 1, avec cette différence qu'ici nous utiliserons r et $s, \dots$ au lieu des n .

b3) Nous calculons r_1 et $s_1, \dots$ valeurs optimum de r et $s, \dots$ quand on tient compte des erreurs d'arrondissement.

Mais cette fois nous fixons le coût de sondage égal à c_0 valeur que nous avons déjà calculé au cours de l'étape b1; et nous calculons r et $s, \dots$ telles que $V = V_1 + V_2$ soit minimum. Ce calcul qui est le but principal des problèmes du type 2, nous permet de calculer les valeurs optimum de r et s , compte tenu des erreurs d'arrondissement, et en même temps nous donne la valeur $V = V_1 + V_2$ de la variance de notre estimateur. Il est évident que

$$V_0 < V < V'_0 \quad (4)$$

où V_0 a été calculée en b1 et V'_0 au cours de b2.

Pour résoudre chaque problème concernant un plan spécial de sondage, nous suivrons, au cours de cette partie, les étapes mentionnés ci-dessus sans reprendre les explications déjà données.

Pour terminer ce paragraphe, disons un mot de l'enquête pilote de Téhéran et Damavand dont les données seront utilisées, au cours de cette partie et de la suivante, pour les applications numériques de la plupart des plans étudiés.

En 1949, nous avons procédé à une enquête-pilote dans les villages de deux provinces de l'Iran, nommées Téhéran et Damavand. Le rapport de cette enquête a été publié par le gouvernement, en persan. Une des rubriques étudiées par cette enquête était la superficie de terre cultivée, annuellement, par chaque

paysan, en blé et en orge.

Nous désignerons, ci-après, cette enquête comme "l'enquête TD".

2- Fonction de coût-

Nous utiliserons plusieurs fois au cours de cette thèse, des fonctions de coût. Parmi les fonctions de coût déjà utilisées par divers auteurs, il y en a quelques-unes qui sont assez exactes pour nous permettre d'étudier l'effet des grandes variations des effectifs de sondage et des tailles des unités de sondage. Ces fonctions ont des structures mathématiques plus ou moins compliquées. Par exemple, les recherches de Mahalanobis et autres ont montré que, pour étudier les variations des effectifs de sondage à deux degrés, un terme contenant $\sqrt{n}$, où n est le nombre des PSU soumises à l'enquête, améliore beaucoup la fonction de coût. Mais ces sortes de fonctions de coût relativement compliquées, peuvent être utilisées, seulement, dans les cas où les fonctions de variance sont assez simples pour que la solution des équations qu'on obtient par les méthodes de l'optimum, par exemple par la méthode $V + \lambda C_{\text{minimum}}$, soit pratiquement possible.

Les fonctions de variance obtenues au cours de la première partie sont la somme de deux termes $V_1 + V_2$, et on sait que les fonctions de coût ont été établies pour être étudiées, avec des fonctions de variance qui ne sont autres que nos V_1 , elles-mêmes beaucoup plus simples que les V_2 (car les V_2 contiennent toujours des termes de degré plus élevé en r , s , etc.. que les V_1).

$V_1 + V_2$ est donc encore plus compliqué. On voit que les fonctions de coût généralement utilisées avec V_1 ne peuvent pas être utilisées avec $V_1 + V_2$. Nous sommes donc obligés de choisir pour nos études des fonctions de coût assez simples.

Celles que nous avons choisies sont basées sur la fonction

$$c = an + b S_i n_i \quad (17)$$

fonction de coût d'un sondage à deux degrés, où a est le coût général de l'enquête pour une PSU moyenne, et b est celui pour un élément dans une PSU moyenne.

En choisissant cette fonction et son mode d'emploi, nous avons fait les approximations suivantes:

1) En réalité, nous devrions avoir comme deuxième terme $S_i n_i b_i$ et non pas $b S_i n_i$, car le coût de l'enquête d'un élément varie d'une PSU à l'autre. Mais nous avons utilisé l'approximation $b_i = b$ pour simplifier les calculs. Cette approximation est la même que celle que fait Neyman pour obtenir la formule de répartition optimum de l'effectif total entre les strates.

2) Peut-être la fonction $a \sqrt{n}$, ou quelque autre, serait-elle plus exacte comme premier terme de (1). Mais ces formes plus compliquées que (an) rendent la solution des équations optimum, si-non impossible, du moins difficile.

Cependant l'approximation (an) semble ne pas être très loin de la vérité, au moins pour les pays sous-développés où le réseau routier est si mauvais que, pour aller d'une PSU à l'autre, on est souvent obligé de retourner au centre de la strate, et

aller de là à l'autre PSU. C'est le cas que l'auteur a constaté pour beaucoup de villages de l'Iran.

En France on est arrivé au même résultat, mais pour une cause un peu différente. Car en France tous les déplacements se font par chemin de fer; et les enquêteurs ne sont pas libres d'aller d'une unité de sondage à l'autre directement (voir P. Thionet, Ref. 4).

Enfin nous devons avouer le caractère un peu fallacieux de cette fonction pour le cas des grandes variations de la taille et de la structure de l'échantillon.

3- La troisième approximation que nous avons faite pour établir la fonction (1) concerne le cas où nous réunissons plusieurs (par exemple q en moyenne) des PSU naturelles, pour construire une PSU nouvelle plus grande partie. Nous changeons dans ce cas la fonction (1) en

$$c = aqn + bq \sum_i S_i n_i \quad (18)$$

Les raisons que nous pensons avoir pour faire cette hypothèse sont:

3a) Si nous laissons n et n_i fixes, il est clair que le nombre des éléments, dans les PSU que nous allons soumettre à l'enquête devient q fois plus grand. C'est pourquoi nous avons changé $b \sum_i S_i n_i$ de la formule (17) en $bq \sum_i S_i n_i$ dans la formule (18). Mais il faut noter que le coût de l'enquête dans les grandes PSU devient plus grand que q fois, car les frais de déplacement sont beaucoup plus grands que pour des petites PSU. D'autre part,

cette augmentation de coût, au moins pour les pays sous-développés (pour la raison que nous avons donnée ci-dessus) est plutôt de même nature que a. Nous l'avons donc ajoutée à a, et non pas à b.

3b) Si la différence de frais de déplacement dans les PSU n'existait pas, nous aurions laissé le premier terme de (17) sans modification quand nous sommes passé de (17) à (18). Mais nous avons vu à l'alinéa (3a) que les frais de déplacement augmentent avec q. Les frais généraux de chaque PSU choisie augmentent aussi. Nous avons tenu compte de ces deux augmentations en changeant (an) de la formule (17) en (aqn) dans la formule (18). Mais nous n'avons pas donné de raisons assez probantes pour la forme de cette transformation. Nous admettons donc que cette hypothèse n'est bonne que pour des petites variations de q, et peut-être fallacieuse pour de grandes variations de q.

Enfin il semble évident que la fonction de coût doit être telle que le coût soit compris entre

$$C = aqn + bq S_i n_i \quad (18)$$

et

$$C = an + bq S_i n_i \quad (19)$$

Nous allons donc utiliser à la fois ces deux formules dans nos calculs, ce qui est très simple dans le cas des calculs mathématiques; (car nous pouvons d'abord faire les calculs avec la formule (18) que nous avons admise de préférence; et puis au lieu de recommencer les calculs avec la formule (19), nous divisons simplement chaque a par q pour obtenir les résultats du calcul

avec la formule (19). Dans le cas des calculs numériques, le procédé n'est plus aussi simple, et nous serons obligés de calculer avec chaque formule (18 et 19), séparément.

Pour chaque plan nous donnerons une forme convenable de la formule (17); et nous donnerons, seulement, de nouvelles explications si elles semblent nécessaires.

3-Répartition optimum au point de vue c constant-

3.1.-Le " problème de Neyman " pour un simple plan de sondage stratifié-

Afin de rendre minimum la variance dans le cas d'une stratification, on répartit l'effectif total n de l'échantillon entre les strates, d'après la formule de Neyman suivant les cinq étapes de calcul:

a₁) Ici les formules classiques sont:

$$n_i = \frac{X_i \sigma_i n}{\sum X_i \sigma_i} \quad (20)$$

où

$$X_i = \sqrt{N_i / (N_i - 1)}$$

N_i nombre des éléments de la strate n°i, n_i = effectif de l'échantillon de la strate n° i et σ_i = écart-type de la strate n°i.

$$V_{opt} = \frac{1}{n} \left(\sum X_i \sigma_i^2 \right) - \sum \frac{N_i \sigma_i^2}{N_i - 1} \quad (21)$$

a2) Comme nous l'avons mentionné au 1, ordinairement, on calcule d'abord les valeurs théoriques des n_i , qui ne sont pas

nécessairement entières. Nous devons donc les arrondir.

Après avoir calculé les n_i par la formule (20), nous choisissons comme effectif de l'échantillon de la strate $n^{\circ}i$ un entier tel que :

$$E(\mu_i) = n_i, \quad E\left(\frac{1}{\mu_i}\right) \neq \frac{1}{n_i} + \frac{V_{m_i}}{n_i^2} \quad (22)$$

(Voir § 8 de la première partie).

a3) D'après les formules (20), (21) et (22)

$$\begin{aligned} V[\hat{X}(P)] &= \sum_i x_i^2 \sigma_i^2 \left(\frac{1}{n_i} + \frac{V_{m_i}}{n_i^2} \right) - V' \\ &= \frac{V_m}{n^2} \left(\sum x_i \sigma_i \right)^2 \sum \frac{1}{x_i \sigma_i} - V' \\ &\quad + \frac{1}{n} \left(\sum x_i \sigma_i \right)^2 \\ &= V_{opt} + \frac{V_m}{n^2} \left(\sum x_i \sigma_i \right)^2 \sum \frac{1}{x_i \sigma_i} \end{aligned}$$

où V_{opt} = la variance minimum théorique de la formule (21). On voit que $V[\hat{X}(P)]$ comprend deux termes, dont le premier est la variance minimum théorique V_{opt} et le deuxième une quantité positive qui s'ajoute à V_{opt} , à cause de la non-divisibilité des N_i par les n_i . Si nous négligeons le terme $-V'$ de V_{opt} (correction de population finie) et si nous posons:

N = nombre de strates,

Ma = moyenne arithmétique des $x_i \sigma_i$

Mh = moyenne harmonique des $x_i \sigma_i$

$\bar{n}$ = moyenne arithmétique des n_i ,

il vient:

$$V_{opt} = \frac{1}{n} \left(\sum x_i \sigma_i \right)^2$$

$$V[\hat{X}(r)] = V_{opt} \left[1 + \frac{V_{m1}}{n^3} \frac{M_a}{M_R} \right]$$

et nous posons

$$t = 1 + \frac{V_{m1}}{n^3} \frac{M_a}{M_R} \quad (23)$$

Ici t ne contient pas V et est donc bien une fonction non-décroissante de V . On vérifie de même que t est toujours plus grand que 1, donc que

$$V[\hat{X}(r)] > V_{opt}$$

On sait que M_a/M_h est toujours supérieur (ou égal à 1) et peut-être théoriquement aussi grand que possible. Mais on peut constater sur des plans de sondage réels que ce rapport n'est jamais supérieur à 2 ou 3. Suivent cinq exemples de ce fait.

Notre premier exemple est emprunté à Cochran (voir p. 35 de Ref.5).

SUPERFICIE MOYENNE DE LA TERRE CULTIVEE EN MAIS
JEFFERSON COUNTRY, IOWA.

Numéros de Strates	Superficie des Fermes en Acres	Nombre de Fermes N_i	G_i
1	0-40	394	8,3
2	41-80	461	13,3
3	81-120	391	15,1
4	121-160	334	19,8
5	161-200	169	24,5
6	201-240	113	26,0
7	241-	148	35,2

$$M_a / M_h = 109$$

Notre deuxième exemple numérique est emprunté à l'ouvrage de Sukhatme (voir page 126 de Ref. 6).

EXPERIENCE DE RECOLTE DU RIZ A CEYLAN 1951-1952

Numéros de Strates	N_i	G_i	
1	189	66	
2	242	122	chaque strate est
3	146	138	formée par un dis-
4	178	41	trict.
5	287	59	

$$M_a / M_h = 1,24$$

Notre troisième exemple est emprunté à l'ouvrage de Yates (voir Ref. 7, pages 203 et 151).

FERMES DU HETFORDSHIRE, 1939

SUPERFICIE EN BLE DES FERMES

Numéros de Strates	Superficie des Fermes (en Acres)	Nombre de fermes N_i	G_i
1	1-5	435	-
2	6-20	519	1,4
3	21-50	357	3,4
4	51-150	519	12,7

Numéros de Strates	Superficie des Fermes (en Acres)	Nombre de fermes	
		N_i	σ_i
5	151-300	400	27,6
6	301 et plus	266	43,9

Pour la strate n°1, σ_i n'a pas été indiqué. Nous avons donc calculé pour les 5 dernières strates:

$$M_a / M_h = 3,17$$

Notre quatrième exemple est emprunté à l'enquête de Fars (voir Ref. 8 page 11).

RECENSEMENT DE 1953: NOMBRE DES OVIDES PAR VILLAGE

Numéros de Strates	Nombre d'habitants par village	Nombre de villages	
		N_i	σ_i
1	0	236	325
2	1-50	482	231
3	51-100	544	359
4	101-200	769	330
5	201-500	722	632
6	501-1000	208	906
7	1001-2500	104	1830

La strate 8, composée de 24 villages dont le nombre d'habitants par villages est supérieur à 2500, sera énumérée. Ici:

$$M_a / M_h = 1,33$$

Notre cinquième exemple est fourni par " l'enquête TD" mentionnée à la fin du §1.

ENQUETE PILOTE 1948: SUPERFICIE CULTIVEE DES VILLAGES

Numéros de Strates	Nombre d'Hab. par village	N_i	σ_i
1	1-50	130	36
2	51-150	388	34
3	151-250	211	52
4	251-350	131	63
5	351-550	165	87
6	551-850	101	74
7	851-1150	32	78
8	1851-3150	29	149

La strate 9, composée de 13 villages dont le nombre d'habitants par village est supérieur à 3150, sera énumérée. Ici on a:

$$M_a / M_h = 1,49$$

Les valeurs obtenues pour M_a / M_h , d'après les exemples ci-dessus, sont 1,09, 1,24, 3,17, 1,33 et 1,49. Sans pouvoir le démontrer mathématiquement, d'après ces expériences numériques, mon opinion est que, dans la plupart des cas, M_a / M_h est compris entre 1 et 2. Donc, si on ne veut pas (ou si en raison de la non-disponibilité des données, on ne peut pas) calculer M_a / M_h , on peut le prendre égal à deux.

Pour compléter nos calculs, nous donnons la table suivante:

COMPARAISON DES RESULTATS DES 5 EXEMPLES CI-DESSUS POUR

UNE STRATIFICATION AVEC SONDAGE AU 20eme et $V_m = \frac{1}{12}$

Exemple	Nombre des Strates	$N. = \sum N_i$	M_a / M_h	$d = t-1$
Cochran	7	2010	1,09	0,00041
Sukhatme	5	1042	1,24	0,00095
Yates	5	2061	3,17	0,00060
Fars	7	3065	1,33	0,00023
Téhéran et Damavand	8	1187	1,49	0,00230

On voit que, dans tous les cas, le biais causé par la non-divisibilité des N_i par n_i est négligeable.

3.2- Le " problème de Neyman " pour un simple plan de sondage stratifié, tirage bernouillien et probabilités inégales dans chaque strate.

La population est la même que celle du plan précédent. On se propose de soumettre à l'enquête dans la strate n^0_k , n_k éléments obtenus par tirage bernouillien avec probabilités inégales et tel que

$$\sum_k n_k = n$$

La probabilité p_{ki} de tirer l'élément n^0_{ki} sera telle que

$$\sum_i p_{ki} = 1$$

séparément pour chaque strate K

a1) On démontre que, si on pose

$$\sigma_K^2 = \sum_{k=1}^{N_K} \frac{1}{n_K} \left[\frac{x_{Kk}}{N_K} - \bar{x}_{K0} \right]^2 \quad (24)$$

La formule pour calculer n_K est:

$$n_K = \frac{N_K \sigma_K n}{\sum N_K \sigma_K} \quad (25)$$

et la valeur minimum de la variance, dans ce cas, est:

$$V_{\text{inf}} = \frac{1}{n} \left(\sum N_K \sigma_K \right)^2 \quad (26)$$

a2) On sait qu'il n'y a aucune évidence logique pour que tous les n_K soient des entiers. Nous sommes obligés, si nous voulons tenir compte des erreurs d'arrondissement, d'utiliser une de nos méthodes d'arrondissement, pour remplacer les n_K par des n'_K .

$$a3) \quad V[\hat{X}(\mu)] = E \left[\sum_K N_K \frac{\sigma_K^2}{n'_K} \right] = \sum_K N_K \sigma_K^2 E \left(\frac{1}{n'_K} \right)$$

et en utilisant la formule (20) de la première partie, nous avons

$$\begin{aligned} V[\hat{X}(\mu)] &= \sum_K N_K \sigma_K^2 \left(\frac{1}{n_K} + \frac{V_{\text{inf}}}{n_K^3} \right) \\ &= \sum_K N_K \frac{\sigma_K^2}{n_K} + V_{\text{inf}} \sum_K N_K \sigma_K^2 / n_K^3 \end{aligned}$$

Substituons la formule (25) dans la précédente, nous avons:

$$\begin{aligned}
 V[\hat{X}(H)] &= V_0 \tau t \\
 &+ V_m \sum N_K^2 \sigma_K^2 \left(\sum N_K \sigma_K \right)^3 / N_K^3 \sigma_K^3 \tau^3 \\
 &= V_0 \tau t + V_m \frac{(\sum N_K \sigma_K)^2}{n^2} \frac{(\sum N_K \sigma_K)}{n^2} \sum \frac{1}{N_K \sigma_K} \\
 &= V_0 \tau t + V_m V_0 \tau t \frac{N^2}{n^2} \frac{M_a}{M'_h} \\
 &= V_0 \tau t \left[1 + \frac{V_m}{n^2} \frac{M_a}{M'_h} \right]
 \end{aligned}
 \tag{27}$$

où, comme dans le cas de la répartition optimum avec probabilités égales formule (23), $\bar{n} = n/N$ et M'_a et M'_h sont, respectivement la moyenne arithmétique et la moyenne harmonique de la variable

$N_K \sigma_K$, où σ_K est calculé d'après la formule (24).

Ici aussi, pour montrer l'augmentation relative de la variance causée par les erreurs d'arrondissement, nous calculons:

$$d = t - 1 = \frac{V_m}{n^2} \frac{M_a}{M'_h} \tag{28}$$

Les résultats des exemples numériques, sont ici les mêmes que pour le plan précédent.

4- Répartition optimum au point de vue V constante.-

4.1- L'équivalent du " problème de Neyman " avec un échantillon non plus stratifié mais à deux degrés.

Le plan ici est le même que le plan II de la deuxième partie

où les effectifs de sondage sont n (premier degré) et n_i (deuxième degré). Voici les 5 étapes de calcul:

a1) Un estimateur sans biais du total

$$X = \sum_i \sum_j x_{ij} \quad \text{est:}$$

$$\hat{X} = \frac{N}{n} \sum_i \frac{N_i}{n_i} \sum_j x_{ij}$$

dont la variance est:

$$V[\hat{X}(n)] = \frac{N}{n} \sum_i \left[\frac{x_i^2 \sigma_i^2}{n_i} - \frac{x_i^2 \sigma_i^2}{N_i} \right] + X^2 \sigma_X^2 \left(\frac{1}{n} - \frac{1}{N} \right) \quad (29)$$

où $x_i^2 = \frac{N_i^3}{N_i - 1}$, $X^2 = \frac{N^3}{N - 1}$
entre les éléments de la PSU $n^o i$ et
les totaux X_i , des PSU.

σ_i^2 = variance
 σ_X^2 = la variance entre

La fonction de coût la plus simple et en même temps la plus naturelle est

$$C = an + b \sum_i n_i$$

Pour trouver les valeurs optimum de n et des n_i , on rend minimum l'expression $V + \lambda C$ pour la valeur désirée de V et les variations de n et des n_i . Le résultat de ce calcul est:

$$n_i = x_i \sigma_i / \sqrt{\frac{b}{a n} (V + X^2 \sigma_X^2 / N)} \quad (30)$$

$$n = \left[\sqrt{\frac{b N}{a}} \left(V + \frac{X^2 \sigma_X^2}{N} \right) \sum_i x_i \sigma_i - N \sum_i x_i \sigma_i / N_i + X^2 \sigma_X^2 \right] / \left(V + \frac{X^2 \sigma_X^2}{N} \right) \quad (31)$$

a2) Rien à Signaler

a3)

$$\begin{aligned}
 V[\hat{X}(\mu)] &= E\{V[\hat{X}(\mu)]\} \\
 &= E\left[\frac{N}{\mu} \sum \left(\frac{\bar{x}_i^2 \sigma_i^2}{n_i} - \frac{\bar{x}_i^2 \sigma_i^2}{N} \right) \right. \\
 &\quad \left. + \bar{x}^2 \sigma_x^2 \left(\frac{1}{\mu} - \frac{1}{N} \right) \pm \bar{x}^2 \sigma_x^2 \left[E\left(\frac{1}{\mu}\right) - \frac{1}{N} \right] \right. \\
 &\quad \left. + N E\left(\frac{1}{\mu}\right) \sum \bar{x}_i^2 \sigma_i^2 \left[E\left(\frac{1}{n_i}\right) - \frac{1}{N} \right] \right]
 \end{aligned}$$

Ce qui, d'après la formule (20) de la première partie est égale à :

$$\begin{aligned}
 V[\hat{X}(\mu)] &= \left\{ \frac{N}{n} \sum \bar{x}_i^2 \sigma_i^2 \left(\frac{1}{n_i} - \frac{1}{N} \right) + \bar{x}^2 \sigma_x^2 \left(\frac{1}{n} - \frac{1}{N} \right) \right\} \\
 &\quad + \left\{ \frac{NV_m}{n^3} \left(\sum \frac{\bar{x}_i^2 \sigma_i^2}{n_i} - \sum \frac{\bar{x}_i^2 \sigma_i^2}{N} \right) + \frac{\bar{x}^2 \sigma_x^2 V_m}{n^3} \right\} \\
 &\quad + \left\{ \frac{NV_m}{n^3} \sum \frac{\bar{x}_i^2 \sigma_i^2}{n_i^3} \right\} \\
 &\quad + \left\{ \frac{NV_m}{n} \sum \frac{\bar{x}_i^2 \sigma_i^2}{n_i^3} \right\}
 \end{aligned}$$

Le second membre de la dernière formule a quatre termes dont chacun est écrit entre accolades. Le troisième terme, qui est beaucoup plus petit que les autres, est négligeable. Nous allons donc le supprimer et écrire les autres termes en fonction de V

et n en utilisant les formules (29) et (31).

$$V[\hat{X}(\mu)] = V + \frac{V_m}{n^2} \left(V + \frac{\bar{x}^2 \sigma_x^2}{N} \right) + \frac{V_m}{n} \frac{1}{n} \left(V + \frac{\bar{x}^2 \sigma_x^2}{N} \right) \sqrt{\frac{b}{a n}} \left(V + \frac{\bar{x}^2 \sigma_x^2}{N} \right) \sum \frac{1}{n_i^3}$$

En négligeant le terme $\sum \sigma_x^2 / N$ (terme pour correction de population finie), nous aurons

$$V[\hat{X}(M)] = tV \quad (32)$$

où

$$t = 1 + \frac{V_m}{n_2} + \frac{V_m b \sqrt{\frac{b}{a}} \frac{V}{N} \sum \frac{1}{N_k}}{a n}$$

donc $t > 1$ toujours.

V = valeur de variance que nous désirons avoir; et $V[\hat{X}(M)]$ est la valeur que la variance aura parce que n et n_1 ne sont pas des entiers.

Application Numérique

Pour faire une application numérique des résultats théoriques de ce §, nous allons admettre :

$$N-1 \neq N, N_1-1 \neq N_1 \text{ et } \sigma_c \neq \sigma$$

Le coefficient de la formule (32) devient:

$$t = 1 + \frac{V_m}{n_2} + \frac{V_m b \sqrt{\frac{b}{a}} \frac{V}{N} \sum \frac{1}{N_k}}{a n \sigma}$$

Ici nous allons appliquer notre résultat aux données de "l'enquête TD". Les données nécessaires pour le présent plan sont: $N = 1281$; $N_1 = N_2 = 72000$; $\sigma_c^2 = 0,8$;

$$\sigma_x^2 = 6950 ; \quad \sum 1/N_k = 23 ; \quad b/a = 0,18 ;$$

$X = 80.000$ hectares de céréales.

Nous allons prendre $V_0 = 64.000.000$ pour que le coefficient de variation soit 10% . Les résultats du calcul où nous avons utilisé la méthode par itération mentionnée au § 1, sont donnés dans le tableau suivant:

CALCUL DE LA VALEUR V_1 QU'ON DOIT DONNER A V POUR
QUE LA VARIANCE TOTALE SOIT 64.000.000-

V_i	t_i	$V_i =$ variance totale
64.000.000	1,141	73.000.000
56.000.000	1,120	62.700.000
57.100.000	1,124	64.300.000
56.900.000	1,125	63.950.000

Sans avoir besoin de continuer nos calculs, nous pouvons prendre:

$$V_5 = \frac{56.900.000 + 57.100.000}{2} = 57.000.000$$

Valeur assez bonne pour le calcul approché que nous faisons.

4.2 - Extension pour K degrés de sondage-

Nous allons généraliser, au cours de ce § l'étude du paragraphe précédent. Nous allons prendre une population divisée en N PSU dont nous soumettons à l'enquête (n) désignées par tirage exhaustif et probabilités égales. La PSU $n^o i$ est divisée en N_i unités du 2^o degré de sondage, dont n_i seront soumises à l'enquête ; et ainsi de suite, nous continuons la division des unités de chaque degré de sondage aux unités plus petites. En

général, nous divisons l'unité $n^0 i_1 i_2 \dots i_k$ de la division de degré α en $N_{i_1 i_2 \dots i_k}$ unités de degré $\alpha + 1$ dont nous soumettons, par tirage au sort, $n_{i_1 i_2 \dots i_k}$ à l'enquête ; et nous continuons ce procédé jusqu'au degré K ; c'est-à-dire nous divisons les unités $n^0 i_1 i_2 \dots i_{k-1}$ de la division $(K-1)$ en $N_{i_1 i_2 \dots i_{k-1}}$ unités d'ordre k dont $n_{i_1 i_2 \dots i_k}$ seront désignées par tirage au sort et mesurées.

a1) On sait que si nous désignons $i_1 i_2 \dots i_k$ par α (le même α que nous avons utilisé au § 8 de la première partie et au § 1 de la deuxième partie) et si nous désignons la mesure de l'unité $n^0 k$ (où k désigne $i_1 i_2 \dots i_k$) par x_k , un estimateur sans biais de X (total de la population) est:

$$\hat{X}(n) = \frac{N}{n} \sum_1 \frac{N_1}{n_1} \sum_2 \dots \sum_{K-1} \frac{N_{K-1}}{n_{K-1}} \sum_K x_K \quad (33)$$

dont la variance est:

$$V[\hat{X}(n)] = \frac{N}{n} \sum_1 \frac{N_1}{n_1} \sum_2 \dots \sum_{K-1} \frac{N_{K-1}}{n_{K-1}} \sum_{K-1} \sigma_{K-1}^2 x_{K-1}^2 \left(\frac{1}{n_{K-1}} - \frac{1}{N_{K-1}} \right) + \dots + \frac{N}{n} \sum_K \sigma_K^2 x_K^2 \left(\frac{1}{n_K} - \frac{1}{N_K} \right) + x_K^2 \left(\frac{1}{n} - \frac{1}{N} \right)$$

où σ_K désigne $\sqrt{\frac{N_K}{N_K - 1}}$

Pour simplifier les formules, nous allons utiliser les notations suivantes:

$$A_{\alpha n} = \frac{N}{n} \sum_1 \frac{N_{\alpha-1}}{n_{\alpha-1}} \sum_1 \frac{x_{\alpha}^2 \sigma_{\alpha}^2}{n_{\alpha}} \quad (\text{ou } A_{\alpha} \quad \text{S'il n'y a pas de confusion à craindre})$$

$$A'_{\alpha} = \frac{N}{n} \sum_1 \dots \frac{N_{\alpha-1}}{n_{\alpha-1}} \sum_{\alpha} \frac{x_{\alpha}^2 \sigma_{\alpha}^2}{N_{\alpha}}$$

$$A_{\alpha}(z) = \frac{N}{n} \sum_1 \dots \frac{N_{\alpha-1}}{n_{\alpha-1}} \sum_z \frac{N_{\alpha}}{n_{\alpha}}$$

$$A'_{\alpha}(z) = \frac{N}{n} \sum_1 \dots \frac{N_{\alpha-1}}{n_{\alpha-1}} \sum_z \frac{N_{\alpha}}{n_{\alpha}} \sum_{\alpha} \dots \sum_{\alpha} \frac{x_{\alpha}^2 \sigma_{\alpha}^2}{N_{\alpha}}$$

Nous avons donc:

$$V[\hat{X}(1)] = A_0 n - A'_0 + A_1 n - A'_1 + \dots + A_{(K-1)} n - A'_{(K-1)} n \quad (34)$$

Malheureusement nous ne sommes pas en mesure de généraliser, exactement, toutes les opérations mathématiques du paragraphe précédent, car si nous prenons même la fonction de coût la plus simple, c'est-à-dire

$$C = a_0 n + a_1 \sum_1 \eta_1 + \dots + \sum_{K-1} \eta_{K-1} \quad (35)$$

nous ne savons pas résoudre le système des équations que nous obtenons, si nous rendons minimum l'expression $V + \lambda C$, comme nous l'avons fait pour le cas de $K = 2$ au paragraphe précédent, même pour $K = 3$.

Cependant, comme la sorte de problèmes que nous étudions ici est l'étude des erreurs d'arrondissement et que la solution du problème énoncé ci-dessus n'est pas notre but, nous allons d'abord supposer que la solution de ce problème est connue; ou mieux encore, nous supposons que nous connaissons une loi (dépendant de la valeur V que nous voulons imposer à la formule (34), d'après laquelle nous pouvons calculer les valeurs n_{λ} . Il est clair que cette loi est constituée par quelques formules de même nature que les formules (30) et (31). Puis, nous allons étudier la partie de la variance ainsi calculée, qui est causée par le remplacement des valeurs théoriques des n_{λ} (qui ne sont pas nécessairement entières) par des entiers.

Toute cette étude sera faite d'abord d'une manière générale, puis nous allons appliquer le résultat que nous obtenons à une solution spéciale de l'équation $V + \lambda C = \text{minimum}$, pour le cas où $K = 3$.

- Cas général du sondage à K degrés-

a2) Après avoir calculé les valeurs non-entières théoriques des n_{λ} d'après la loi mentionnée en (a1), nous pouvons choisir les entiers qui doivent les remplacer d'après l'une ou l'autre des méthodes d'arrondissement.

a3) En utilisant les notations données au paragraphe a₁ et en remplaçant les n de la formule (34) par μ , nous avons:

$$V[\hat{X}(\mu)] = E[A_{0\mu} - A'_0 + A_{1\mu} - A'_1 + \dots + A_{(K-1)\mu} - A'_{K-1}]$$

Et en utilisant la formule (20) de la première partie, pour $E(1/\mu)$, nous avons:

$$\begin{aligned}
 V[\hat{\chi}(\mu)] = & V + V_m [A_{K-1}(0) - A'_{K-1}(0) + \dots \\
 & + A_1(0) - A'_1(0) + A_0(0)] \\
 & + V_{m,1} [A_{K-1}(1) - A'_{K-1}(1) + A_2(1) - A'_2(1) + A_1(1)] \\
 & + \dots \\
 & + V_m [A_{K-1}(K-2) - A'_{K-1}(K-2) + A_{K-2}(K-2)] \\
 & + V_m [A_{K-1}(K-1)] \quad (36)
 \end{aligned}$$

Le nombre des termes $A_{\chi}(r)$ de la formule (35) est

$$(k+1)K/2,$$

et le nombre de ses termes $-A'_{\chi}(r)$, qui sont les corrections pour population finie, est $k(k-1)/2$, car dans chaque accolade le nombre des $A'_{\chi}(r)$ est un de moins que celui des $A_{\chi}(r)$

Lorsque K est petit, on peut calculer facilement chaque terme de la formule (35). Mais quand k est grand, en considérant que dans la pratique la somme de tous les termes de la formule (35) est d'ordre plus petit que V , nous pouvons procéder à l'approximation suivante:

(1)- Dans chacun des termes $A_{\chi}(r)$ et $A'_{\chi}(r)$ d'après les définitions que nous avons données, le nombre n_r a l'exposant 3, tandis que les autres n_{χ} ont l'exposant 1. Pour pouvoir

exprimer ces termes en termes des A_{α} et A'_{α} , nous transformons n_{α}^3 en $(n_{\alpha}) (\bar{n}_{\alpha}^2)$ ou $\bar{n}_{\alpha}$ est la moyenne arithmétique de tous les n_{α} . Cette approximation semble raisonnable, car à cause de la grandeur de k , les erreurs que nous aurons pour chaque n_{α} seront compensées les unes par les autres d'après la loi des erreurs.

Nous aurons donc:

$$\begin{aligned} V[\hat{X}(\mu)] = & \frac{V_m}{n^2} (A_{k-1} - A'_{k-1} + \dots + A_0) \\ & + \frac{V_m}{n_1^2} (A_{k-1} - A'_{k-1} + \dots + A_1) \\ & + \dots \dots \dots \\ & + \frac{V_m}{n_{k-1}^2} A_{k-1} \end{aligned} \quad (37)$$

La solution numérique de la formule (37) est beaucoup plus facile que celle de la formule (36).

(2)- Si on veut utiliser une formule encore plus approximative, mais plus simple et facile à manier, on peut utiliser les deux principes suivants:

- Les valeurs des A'_{α} sont beaucoup plus petites que celles des A_{α} . On peut donc négliger leur présence ou absence quand on en a besoin et,

- Les valeurs des A_{α} sont indépendantes les unes des autres, à cela près que leur somme doit être égale à V . Nous pouvons donc prendre chacune d'elles en moyenne égale à V/K ,

ce qui transforme la formule (37) en

$$V[\hat{X}(\mu)] = V \left[1 + \frac{V_m}{K} \left(\frac{K}{n^2} + \frac{K-1}{n_1^2} + \dots + \frac{1}{n_{K-1}^2} \right) \right] = Vt \quad (38)$$

ou

$$t = 1 + \frac{V_m \sum_{\alpha=1}^{K-1} \frac{(K-\alpha)}{n_\alpha^2}}{K}$$

Au point de vue de la symétrie, la formule (38) a ce défaut que, dans la sommation de t , les A_α avec les plus grands indices α ont plus de poids que les autres. Comme cette formule n'est qu'approchée, ce défaut nous semble être négligeable.

- Application à un cas spécial -

Comme nous l'avons mentionné, on ne sait pas résoudre les équations qui explicitent la condition $V + \lambda c$ minimum. Pour avoir néanmoins un champ d'application objective des résultats théoriques que nous avons obtenus, nous allons résoudre ce problème d'une façon dont le résultat est, plus ou moins, différent du minimum absolu de $V + \lambda c$.

Nous admettons qu'en même temps que $V + \lambda c$ doit être minimum, les n_i (c'est-à-dire les nombres des échantillons prélevés dans les PSU) sont, comme dans le cas de $K = 2$ (§4.1) égaux à

$$n_i = \frac{x_i s_i}{\sqrt{\frac{b}{an}} (V + x^2 c_x^2 / N)} \quad (39)$$

comme dans la formule (30) où $\delta_i = N_i^3 / (N_i - 1)$

En substituant la formule (39) dans la formule (34) pour $K = 3$, et rendant minimum l'expression $V + \lambda C$ où c'est la fonction (35) pour $K = 3$, et en admettant les approximations:

$$\sigma_i^2 \# \sigma_1^2, \sigma_{ij}^2 \# \sigma_2^2, N-1 \# N, N_i-1 \# N_i \text{ et } N_{ij}-1 \# N_{ij}$$

nous aurons:

$$nA_0 = N \sigma_X^2 \quad nA'_0 = N \sigma_X^2$$

$$nA_1 = \sqrt{\frac{b}{a} N (V + N \sigma_X^2)} N_0 \sigma_1 \quad nA'_1 = N_0 \sigma_1^2$$

$$nA_2 = \frac{1}{\sigma_1^2} \sqrt[4]{\frac{b}{a} N (V + N \sigma_X^2)^3} \sqrt{\frac{c}{a}} N_{00} \sigma_2$$

$$nA'_2 = \frac{1}{\sigma_1^2} \sqrt{\frac{b}{a} N (V + N \sigma_X^2)} N_{00} \sigma_2^2$$

$$n_{ij} = \sqrt{\frac{V \frac{b}{a} N}{\frac{c}{a} V + N \sigma_X^2}} N_{ij} \sigma_2$$

$$\eta = \frac{nA_0 + nA_1 + nA_2 - nA'_1 - nA'_2}{V + A'_0}$$

$$\bar{\eta}_1 = \frac{N_0 \sigma_1^2}{\sqrt{\frac{b}{a} N \sqrt{V + N \sigma_X^2}}}$$

$$\bar{\eta}_2 = \frac{\sqrt[4]{\frac{b}{a} N N_{00} \sigma_2}}{\sqrt{\frac{c}{a}} \sqrt[4]{V + N \sigma_X^2} N_0}$$

Appliquons les résultats de ce paragraphe à une exemple numérique.

- Exemple Numérique -

Nous utilisons les données de "l'enquête TD". Quelques-unes des données que nous utilisons ici sont estimées à partir des autres. Nous avons évalué σ_X^2 d'après ce que nous avons de la distribution de X_1 en Iran.

Notre plan est le suivant:

(1) Les villages de l'Iran sont divisés en 1000 subdivisions administratives, chacune contenant en moyenne 42 villages. Ces subdivisions, qui sont les unités les plus naturelles du pays, après les villages eux-mêmes, seront considérées dans notre plan comme des PSU.

(2) De chaque PSU choisie, nous tirons au sort quelques villages, n_1 .

(3) Du village choisi n_{0ij} , nous tirons au sort n_{jk} paysans dont nous mesurons la superficie des terres cultivées en céréales (en hectares).

(4) Notre but est d'estimer la superficie totale des terres du pays cultivées en céréales.

(5) Nous supposons que les résultats trouvés pour la région de Téhéran et Damavand, au point de vue de la variance, sont applicables à tout le pays.

Les données utilisées dans les calculs sont:

$$\begin{aligned} N &= 1000 ; N_0 = 42.000 ; N_{00} = 2.400.000 ; \\ \sigma_1^2 &= 0,8 ; \sigma_2^2 = 3.000 ; \sigma_X^2 = 4.000.000 ; \\ X &= 2.600.000 \text{ hectares} ; b/a = 0,1 ; c/a = 0,01. \end{aligned}$$

Nous avons appliqué les formules trouvées plus haut à ces données et les résultats sont les suivants, pour un coefficient de variation égal à 5% et $V_m = 1/12$.

$$V = 16.900.000.000 ; nA_1 = 3.330.000.000.000$$

$$nA'_1 = 124.000.000.000$$

$$nA_2 = 560.000.000.000$$

$$nA'_\Delta = 47.000.000.000$$

$$nA_0 = 4.000.000.000.000$$

$$A'_0 = 4.000.000.000$$

$$n = 345$$

$$\bar{n}_1 = 1,5$$

$$\bar{n}_2 = 4,1$$

$$t = 1,0255 \text{ d'après la formule (38)}$$

$$t = 1,0395 \text{ d'après la formule (37)}$$

Comme ici t est très peu différent de 1, nous ne sommes pas obligés d'utiliser la méthode par itération. Il suffit que nous prenions

$$V'_0 = 16.900.000.000/1,0395 = 16.300.000.000$$

4.3- Tirage exhaustif de deux PSU (en tout) et probabilités inégales.

Le plan de ce sondage est le même que le plan V de la deuxième partie. Mais ici nous sommes intéressés à calculer la valeur optimum de n_i , effectif de l'échantillon de la PSU n^0_i .

a1) On adopte ici:

$$C = a + b \sum_{i=1}^N P_i n_i \quad (40)$$

où P_i est, comme nous avons défini au cours du paragraphe 6 de la seconde partie, la probabilité totale pour que la PSU n°i soit dans l'échantillon; et dans le cas où $n = 2$, a est les frais généraux d'enquête de 2 PSU moyennes et b est le coût de l'enquête de deux éléments moyens.

L'estimateur sans biais de ce plan est:

$$\hat{X} = \sum_i \frac{N_i}{P_i n_i} \sum_j x_{ij} \quad (41)$$

et sa variance est:

$$\begin{aligned} V[\hat{X}(n)] &= \sum_i \frac{N_i(N_i - n_i)}{P_i(N_i - 1)} \frac{\sigma_i^2}{n_i} \\ &\quad + \sum_{i \neq i'} \sum_{j \neq j'} x_{ij} x_{i'j'} \left(\frac{P_{ii'}}{P_i P_{i'}} - 1 \right) + \sum_i x_{ii}^2 \left(\frac{1}{P_i} - 1 \right) \\ &= \sum_i \frac{x_{ii} \sigma_i^2}{P_i n_i} + V' \end{aligned} \quad (42)$$

La solution de l'équation $V + \lambda c = \text{minimum}$ est:

$$n_i = \frac{x_{ii} \sigma_i \sum_i x_{ii} \sigma_i}{P_i (V - V')} \quad (43)$$

a2) Comme pour les autres problèmes, nous arrondissons n_i à μ_i d'après une des cinq méthodes.

a3) La variance $V[\hat{X}(\mu)]$ devient:

$$\begin{aligned} V[\hat{X}(\mu)] &= E \left[\sum_i \frac{x_{ii}^2 \sigma_i^2}{P_i} E \left(\frac{1}{\mu_i} \right) + V' \right] \\ &= \sum_i \frac{x_{ii}^2 \sigma_i^2}{P_i} \left(\frac{1}{\mu_i} + \frac{1}{\mu_i^3} \right) + V' \end{aligned}$$

Substituons (43) dans la dernière formule, nous aurons:

$$V[\hat{X}(V)] = V + V_m \left(\frac{V - V'}{\sum \delta_i \sigma_i} \right)^3 \sum \frac{P_i^2}{\delta_i \sigma_i} = tV \quad (44)$$

où

$$t = 1 + \left(1 - \frac{V'}{V}\right) (V - V')^2 \frac{V_m}{(\sum \delta_i \sigma_i)^3} \sum \frac{P_i^2}{\delta_i \sigma_i}$$

est, manifestement, une fonction croissante de V , et plus grande que 1 pour toutes les valeurs positives de V , car $V > V'$,

a4) Nous pouvons calculer l'augmentation relative causée par les erreurs d'arrondissement, par la formule:

$$d = t - 1$$

où t a été défini dans la formule (44).

a5) Afin de calculer la valeur W , on résoud l'équation du troisième degré en W :

$$W t(W) = V_0$$

- Exemple numérique -

Si nous désignons la probabilité de choix de la PSU $n^o i$ par P_i , les formules qui donnent P_i et P_{ij} sont (voir, par exemple Ref. 6):

$$P_i = \frac{r_i (S + 1 - \frac{r_i}{1 - r_i})}{\sum_i \frac{r_i}{1 - r_i}} \quad \text{où} \quad S = \sum_i \frac{r_i}{1 - r_i}$$

$$P_{ij} = \frac{r_i r_j}{1 - r_i} + \frac{1}{1 - r_j} \quad (45)$$

Nous simplifierons les formules (45) pour nos calculs numériques en admettant $P_i \neq 2P_i$

$$P_{ij} \neq 2p_i p_j, \quad P_i \neq N_i/N. \text{ et } \frac{1}{1 - p_i} \neq 1.$$

Nous transformons les formules (45) en :

$$P_i = \frac{2N_i}{N_0}$$

$$P_{ij} = 2N_i N_j / N_0^2$$

En posant $\sigma_i^2 \neq \sigma^2, N_i - 1 \neq N_i$ et $V_m = \frac{1}{12}$

la formule (44) devient:

$$V[\hat{X}(N)] = V + (V - V')^3 / 3(N_0 \sigma^4)$$

où

$$V' = \frac{N_0}{2} \sum N_i \xi_{i0}^2 - \frac{1}{2} \sum X_i^2 - \frac{X^2}{2} - \frac{N N_0 \sigma^2}{2}$$

Les données de " l'enquête TD" nécessaires pour application de la théorie de ce paragraphe sont : $V' = 2.460.000.000$, ce qui montre que, avec le plan présent, nous ne pouvons pas espérer que la variance soit plus petite que 2.460.000.000 ; $N = 72.000$; nous admettons la valeur 3.875.000.000 pour V_0 (c'est un "minimum indépendant" que nous avons défini au § 1, et que nous allons trouver au § 5.4).

En substituant ces valeurs dans la formule $V[\hat{X}(\mu)]$ nous trouvons :

$$t - 1 = 0,02 \quad \text{et} \quad W = 3.800.000.000$$

4.4 - Stratification suivie de deux degrés de sondage;

Tirage de 1 PSU de chaque strate avec Probabilités inégales

Le plan de ce sondage est le même que pour le plan VI de la seconde partie. Mais ici nous cherchons à calculer les n_{ki} , nombre des éléments qui doivent être soumis à l'enquête dans la PSU n^o_{ki} , pour que l'expression $V + \lambda c$ soit minimum.

a) La fonction de coût ici est:

$$c = a + b \sum_K \sum_c \tau_{ki} n_{ki} \quad (47)$$

où a est les frais généraux d'enquête de N PSU moyennes provenant chacune d'une strate, et b est le coût de l'enquête de N éléments moyens.

L'estimateur de ce plan est:

$$\hat{X} = \sum_K \frac{N_{ki}}{n_{ki} \tau_{ki}} \sum_j x_{ij} \quad (48)$$

et sa variance

$$\begin{aligned} V[\hat{X}(n)] &= \sum_K \sum_c \frac{N_{ki}^2 (N_{ki} - n_{ki}) \sigma_{ki}^2}{\tau_{ki}^2 (N_{ki} - 1) n_{ki}} + \sum_K \sum_c \tau_{ki} \left(\frac{x_{ki}^2}{\tau_{ki}} - x_K \right)^2 \\ &= \sum_K \sum_c \frac{x_{ki}^2 \sigma_{ki}^2}{\tau_{ki} n_{ki}} + V' \end{aligned} \quad (49)$$

La solution de l'équation $V + \lambda C = \text{minimum}$ est:

$$n_{ki} = \gamma_{ki} \sigma_{ki} \sum_K \sum_C \gamma_{ki} \sigma_{ki} / \tau_{ki} (V - \bar{V}) \quad (50)$$

a2) Comme les n_{ki} ne sont pas nécessairement entiers, nous les arrondissons par une des cinq méthodes, ce qui augmente la variance.

a3)

$$\begin{aligned} V[\hat{X}(P)] &= E\left[\sum_K \sum_C \frac{\gamma_{ki}^2 \sigma_{ki}^2}{\tau_{ki}^2 P_{ki}}\right] + V \\ &= \sum_K \sum_C \frac{\gamma_{ki}^2 \sigma_{ki}^2}{\tau_{ki}^2} E\left(\frac{1}{P_{ki}}\right) + \bar{V} \\ &= \sum_K \sum_C \frac{\gamma_{ki}^2 \sigma_{ki}^2}{\tau_{ki}^2} \left(\frac{1}{n_{ki}} - \frac{V_m}{n_{ki}^3}\right) + \bar{V} \\ &= \bar{V} + \sum_K \sum_C \frac{\gamma_{ki}^2 \sigma_{ki}^2 \tau_{ki}^3}{\tau_{ki}^2 \gamma_{ki}^3 \sigma_{ki}^3} \left(\frac{V - \bar{V}}{\sum_K \sum_C \gamma_{ki} \sigma_{ki}}\right)^3 V_m \\ &= \bar{V} + V_m \left(\frac{V - \bar{V}}{\sum_K \sum_C \gamma_{ki} \sigma_{ki}}\right)^3 \sum_K \sum_C \frac{\tau_{ki}^2}{\gamma_{ki} \sigma_{ki}} \\ &= tV \end{aligned} \quad (51)$$

où t

$$t = 1 + \left(1 - \frac{\bar{V}}{V}\right) \left(\frac{V - \bar{V}}{\sum_K \sum_C \gamma_{ki} \sigma_{ki}}\right)^2 \frac{V_m}{\left(\sum_K \sum_C \gamma_{ki} \sigma_{ki}\right)^3} \sum_K \sum_C \frac{\tau_{ki}^2}{\gamma_{ki} \sigma_{ki}}$$

est manifestement une fonction croissante de V et plus grande que 1.

- Exemple numérique -

En posant $p_k \# N_k / N_0$, $\sigma_{ki} \# \sigma^2$, $N_k - 1 \# N_k$
 $N_{ki} - 1 \# N_{ki}$ et $V_m = \frac{1}{12}$

nous transformons la formule (51) en :

$$V[\hat{X}(1)] = V + (V - V)^3 / 12 (N_0 \sigma)^4$$

où

$$V' = \sum_K N_{K0} \sum_{K0} N_{K0} \sum_{K0} N_{K0} - \sum_K X_K^2 - \sigma^2 \sum_K N_K N_{K0}$$

Nous appliquons la théorie de ce paragraphe à " l'enquête TD" . En fait, les 1281 villages qui formaient le champ de cette enquête étaient stratifiés en 5 strates géographiques; suivant les données tirées de l'enquête.

Nous désignerons, ci-après, cette forme stratifiée de " l'enquête TD" par " enquête stratifiée TD". Les données nécessaires sont : $V' = 256.000.000$, ce qui montre que, avec le présent plan, nous ne pouvons pas espérer que la variance soit moindre que 256.000.000 ; $N_{..} = 72.000$ et $\sigma^2 = 0,8$.

Nous posons $V_0 = 545.000.000$; c'est un " optimum indépendant" (voir § 4.3). Les résultats des calculs sont:

$$t - 1 = 0,01\% \quad \text{et} \quad W = 544.500.000$$

4.5 - Sondage à deux degrés n PSU par tirage bernoullien et Probabilités inégales-

Nous cherchons à calculer les valeurs optimum des n_i , ou plus exactement celles qui rendent (cV) minimum. Théoriquement on pourrait rendre minimum $V + \lambda c$, mais en raison des difficultés de calcul on y a renoncé et on s'est contenté de rendre c V minimum (voir Sukhatme, Ref. 6, page 363).

a) La fonction de coût est ici:

$$c = a_n + b_n \sum_i h_i n_i \quad (53)$$

où a représente les frais généraux d'enquête pour une PSI moyenne et b est le coût moyen de l'enquête pour un seul élément de la population.

L'estimateur est:

$$\hat{X} = \frac{1}{n} \sum_i \frac{N_i}{h_i n_i} \sum_j x_{ij} \quad (54)$$

et sa variance est:

$$\begin{aligned} V[\hat{X}(n)] &= \frac{1}{n} \sum_i \frac{N_i(N_i - n_i) \sigma_i^2}{h_i(N_i - 1)n_i} + \frac{1}{n} \sum_i h_i \left(\frac{X_i}{h_i} - X \right)^2 \\ &= \frac{1}{n} \sum_i \frac{X_i^2 \sigma_i^2}{h_i n_i} + \frac{1}{n} \left[\sum_i h_i \left(\frac{X_i}{h_i} - X \right)^2 - \sum_i \frac{X_i^2 \sigma_i^2}{h_i N_i} \right] \quad (55) \end{aligned}$$

La solution de l'équation c V = minimum est:

$$n_i = \frac{1}{N_0} \sqrt{\frac{a}{b \Delta}} \frac{X_i \sigma_i}{h_i} \quad (56)$$

où

$$\Delta = \sum f_i \left(\frac{N_i \bar{x}_i^2}{N_0 f_i} - \bar{x}_{00}^2 \right) - \frac{1}{N_0^2} \sum \frac{x_i^2 \sigma_i^2}{f_i N_i}$$

Pour calculer n, on substitue la formule (56) dans (55).

a2) Rien à signaler.

a3)

$$\begin{aligned} V[\hat{X}^{(M)}] &= E \left[\frac{1}{\mu} \sum \frac{x_i^2 \sigma_i^2}{f_i N_i} - \frac{1}{\mu} \sum \frac{x_i^2 \sigma_i^2}{f_i N_i} \right. \\ &\quad \left. + \frac{1}{\mu} \sum f_i \left(\frac{x_i}{f_i} - \bar{x} \right)^2 \right] \\ &= E \left(\frac{1}{\mu} \right) \sum \frac{x_i^2 \sigma_i^2}{f_i} E \left(\frac{1}{N_i} \right) \\ &\quad + E \left(\frac{1}{\mu} \right) \left[\sum f_i \left(\frac{x_i}{f_i} - \bar{x} \right)^2 - \sum \frac{x_i^2 \sigma_i^2}{f_i N_i} \right] \\ &= \left(\frac{1}{n} + \frac{V_m}{n^3} \right) \sum \frac{x_i^2 \sigma_i^2}{f_i} \left(\frac{1}{n_i} + \frac{V_m}{n_i^3} \right) \\ &\quad + \left(\frac{1}{n} + \frac{V_m}{n^3} \right) \left[\sum f_i \left(\frac{x_i}{f_i} - \bar{x} \right)^2 - \sum \frac{x_i^2 \sigma_i^2}{f_i N_i} \right] \\ &= V + \frac{V_m}{n} \sum \frac{x_i^2 \sigma_i^2}{f_i N_i} + \frac{V_m}{n^3} \sum \frac{x_i^2 \sigma_i^2}{f_i N_i} \\ &\quad + \frac{V_m}{n^3} \sum \frac{x_i^2 \sigma_i^2}{f_i N_i} + \frac{1}{n^3} \left[\sum f_i \left(\frac{x_i}{f_i} - \bar{x} \right)^2 - \sum \frac{x_i^2 \sigma_i^2}{f_i N_i} \right] \end{aligned}$$

En négligeant le quatrième terme et considérant que:

$$n = \left[\sum \frac{x_i^2 \sigma_i^2}{f_i N_i} - \sum \frac{x_i^2 \sigma_i^2}{f_i N_i} + \sum f_i \left(\frac{x_i}{f_i} - \bar{x} \right)^2 \right] / V$$

et que n_1 se calcule par la formule (56), nous avons:

$$V[\hat{X}(\mu)] = V \left(1 + \frac{V_m C}{A+B} \right) + V \frac{V_m B}{(A+B)^3} = V t \quad (57)$$

où

$$A = N_0 \sqrt{\frac{L\Delta}{a}} \sum_i x_i \sigma_i^2, \quad B = \sum_i f_i \left(\frac{x_i^2}{f_i} - X^2 \right) - \sum_i \frac{x_i^2 \sigma_i^2}{f_i N_i}$$

$$\text{et } C = \left(N_0 \sqrt{\frac{L\Delta}{a}} \right)^3 \sum_i \frac{f_i^2}{x_i \sigma_i^2}$$

a4) On voit que t a les mêmes propriétés que les t des cas précédents. Afin de calculer l'accroissement relatif de la variance nous calculons

$$d = t - 1$$

a5) Pour calculer W (voir § 1), nous résolvons directement l'équation du troisième degré (57) en V où nous remplaçons V par W et $V[\hat{X}(\mu)]$ par V_0 .

- Application numérique -

Posons $p_i \# N_i/N$, $\sigma_i^2 \# 5^2$, $N_i - 1 \# N_i$ et $V_m = 1/12$, les paramètres de la formule (57) deviennent:

$$\left[\begin{aligned} A &= N_0 \sqrt{\frac{L\Delta}{a}} \\ B &= N_0 \sum_i N_i \sigma_i^2 - X^2 - N_0 5^2 \\ C &= N_0 \left(\sqrt{\frac{L\Delta}{a}} \right)^3 / 5 \end{aligned} \right. \quad (58)$$

En utilisant les données de " l'enquête TD" déjà utilisées au § 4.3, nous avons calculé les paramètres (58):

$$\Delta = 0,986 ; A = 1.750.000.000 ; B = 4.930.000.000 ;$$

$$\text{et } C = 486.000.000$$

Et en substituant ces données dans l'équation (57), nous avons obtenu pour un coefficient de variation 5% ($V_0 = 16.000.000$), les résultats suivants:

$$t - 1 = 0,0061 \quad \text{et} \quad W = 15.830.000$$

5. Les problèmes de type 2 : choix optimum de r. s....

5.1 - Sondage à deux degrés avec probabilités inégales et tirages exhaustifs (plan II)-

Nous avons donné le plan de ce sondage au § 1 de la seconde partie.

bl) Si tous les N_i s'étaient des entiers, la variance de notre estimation serait égale au V_1 de la formule (12) de la seconde partie, et pour choisir les valeurs optimum de s et t, nous devrions rendre minimum l'expression $V_1 + \lambda c$. Nous allons d'abord calculer ces valeurs.

Pour ce calcul, nous avons choisi la fonction de coût

$$c = a/r + b/rs \quad (59)$$

Si nous rendons minimum l'expression $V_1 + \lambda c$, nous obtenons

$$N_0 \sigma_y^2 - \frac{b}{a} (N_0 \sigma_x^2 - N_0 \sigma_y^2) = 0 \quad (60)$$

$$r_0 = \frac{V_1 + N_0 \sigma_x^2}{(s_0 - 1) N_0 \sigma_y^2 + N_0 \sigma_x^2} \quad (61)$$

De (60) on tire s_0 (positif) qu'on porte dans (61) pour obtenir r_0 .

b2)

$$V_c = V_0 + \frac{1}{n} V_m \left[\sum \frac{E_i^2}{E_0^2} - \sum \frac{G_i^2}{n-1} \right]$$

b3) Maintenant nous allons résoudre les deux problèmes que nous avons posés au § 1.

1) Pour tenir compte des erreurs d'arrondissement, nous devons rendre minimum $V + \lambda c$ où $V = V_1 + V_2$ de la formule (12) de la seconde partie et où c est la fonction de coût (59).

Nous obtiendrons.

$$2A\sigma_1^2 + \left(N_0 E^2 + \frac{G}{A}\right) \sigma_1^2 - \frac{G}{A} (N E_X^2 - N_0 E^2) = 0 \quad (62)$$

$$\sigma_1 = \frac{V + N E_X^2}{4\sigma_1^2 + (4-1)N E^2 + N E_X^2} \quad (63)$$

L'équation (62) est cette fois du troisième degré en s_1 ; son terme du premier degré manque et le coefficient de son terme du deuxième degré (qui est toujours positif) est beaucoup plus grand que celui du terme du troisième degré (qui est, lui-même, toujours positif); enfin le terme constant est toujours négatif.

Ces caractéristiques de l'équation (6) nous assurent qu'elle a deux racines négatives et une positive. Nous choisissons, naturellement, la racine positive, comme valeur optimum de s . En substituant cette valeur s_1 de s dans la formule (63), nous obtiendrons la valeur optimum r_1 de r , pour la valeur désirée de la variance.

2) Comme on devait s'y attendre, on vérifie que la valeur donnée pour s par l'équation (60) est toujours plus grande que celle que nous obtenons par l'équation (62); car le terme constant est le même dans les deux, tandis que les coefficients des termes du troisième et deuxième degrés sont, tous deux, plus grands dans (62) que dans (60).

La comparaison entre les formules (61) et (63) n'est pas rigoureusement possible. Mais il semble que, dans la pratique, r_0 calculée par les formules (60) et (61) soit un peu plus petite que r_1 calculée par les formules (62) et (63).

- Exemple numérique -

Nous allons utiliser le résultat de "l'enquête TD" pour le plan d'un sondage à deux degrés, dont le but est d'estimer la superficie totale de blé et d'orge.

Les données numériques que nous empruntons à cette enquête sont:

N_0 = nombre total des paysans = 72.000 ;

N = nombre total de villages = 1.281 ;

$$\sigma^2 = 0,8 ;$$

$$\sum \xi_{i0}^2 = 2.850 ;$$

$$\sigma_{N_c}^2 = 1.200 ;$$

$$\sigma_{\xi}^2 = 1 ; E(\xi^2) = 2,21 ;$$

$$\sum \frac{1}{N_c} = 23 ; \quad b/a = 10 ;$$

X = superficie totalè de terre cultivée, annuellement, en blé et orge = 80,000 hectares.

On voit qu'ici la valeur de N n'est pas suffisante pour espérer obtenir des estimations avec un coefficient de variation petit. Nous basons donc nos calculs sur un coefficient de variation de 5% à obtenir quand on ne tient pas compte des erreurs d'arrondissement. Afin d'obtenir ce coefficient de variation, nous devons avoir $V_0 = 16.000.000$.

Si nous ne tenons pas compte des erreurs d'arrondissement, il vient:

$$s_0 = 39,2 ; r_0 = 2,24 ; c_0 = 0,56a ; V_0 = 16'800'000.$$

Si nous tenons compte des erreurs d'arrondissement, il vient.

$$s_1 = 34,3 ; r_1 = 2,31 ; V = 16'700'000 ; V/V_0\% = 104,4\%$$

Les différences entre ces deux résultats ne sont pas notables

5.2 - Stratification suivie de deux degrés de sondage; tirages exhaustifs et avec probabilités égales (plan III).-

bl) Posons, dans la formule (14) de la seconde partie:

$$V_m \left[\sum_K \sum_{i0}^2 \xi_{K i0}^2 - \sum_K \sum_{i0}^2 \frac{\xi_{K i0}^2}{N_{K i0} - 1} \right] = A$$

$$+ V_m \left[\sum_K \sum_{i0}^2 \xi_{K i0}^2 - \sum_K \frac{\xi_{K i0}^2}{N_K - 1} \right] = B$$

Si nous admettons $N_{k-1} \# N_k$, $N_{ki-1} \# N_{ki}$,
 $\sigma_{ki}^2 \# \sigma^2$, $\sigma_{xk}^2 \# \sigma_x^2$

et posons $N_{00} = \sum_K \sum_i N_{Ki}$, et $N_0 = \sum_K N_K$

nous aurons en première approximation :

$$V_1 = r(s-1)N_{..} \sigma^2 + N_{..}(r-1) \sigma_x^2$$

$$V_2 = A r s^2 + B r^2$$

Si on ne voulait pas tenir compte des erreurs d'arrondissement, ou bien si N_k/r et N_{ki}/s étaient des entiers, quelque soient k et i , pour trouver les valeurs optima de r et s nous devrions rendre minimum l'expression $V_1 + \lambda c$. Nous allons donc d'abord calculer ces valeurs.

Comme fonction de coût nous choisissons la même que pour le plan II, c'est-à-dire;

$$c = a/r + b/rs \quad (64)$$

En rendant minimum l'expression $V_1 + \lambda c$, nous obtenons;

$$N_{..} \sigma^2 s_0^2 - \frac{b}{a} (N_{..} \sigma_x^2 - N_{..} \sigma^2) = 0 \quad (65)$$

$$r_0 = \frac{V_1 + N_{..} \sigma_x^2}{(s_0-1)N_{..} \sigma^2 + N_{..} \sigma_x^2} \quad (66)$$

Les formules (65) et (66) sont analogues aux formules (60) et (61) et possèdent les mêmes propriétés mathématiques.

En substituant r_0 et s_0 dans (64), nous trouvons c_0 :

$$b_2) V_0 = V_0 + A r_0 s_0^2 + B r_0^2$$

b₃) Nous allons résoudre les deux problèmes que nous avons posés au cours du § 1.

Si nous voulons tenir compte des erreurs d'arrondissement, nous devons rendre minimum l'expression $V + \lambda c$. Nous obtenons :

$$2 A s_1^3 + (N.. c^2 + \frac{b}{a} A) s_1^2 - \frac{b}{a} (N.. c^2 - N.. c^2 + 2 B r_1) = 0. \quad (67)$$

L'équation en s_1 (67) est l'analogue de l'équation du troisième degré (62) du § 5.1, avec cette différence qu'ici le terme constant renferme r_1 donc, contrairement au plan II, la valeur minimum de s_1 , ici, dépend de r_1 et de la valeur que nous avons adoptée pour V_0 ou c_0 .

La méthode que nous proposons pour résoudre notre problème consiste à :

- (a) donner différentes valeurs convenables à r ,
- (b) résoudre l'équation (67) pour ces valeurs de r , et trouver les valeurs correspondantes de s ,
- (c) calculer les valeurs correspondantes de c par la formule (64),
- (d) en utilisant les couples de valeurs (r, c) que nous avons ainsi trouvés, calculer par interpolation la valeur r_1 de r qui rend c égal à c_0 obtenu à l'étape (b₁), et enfin

(e) calculer V_1 , V_2 et $V = V_1 + V_2$.

- Application numérique -

Nous appliquons la théorie de ce paragraphe à "l'enquête stratifiée TD" (voir § 4.4).

Si nous ne tenons pas compte des erreurs d'arrondissement, et nous calculons la valeur optimum de r et s pour un coefficient de variation 5%, nous trouvons:

$$s_0 = 24,5 ; \quad r_0 = 3,99 ; \quad \text{et } c_0 = 0,353a.$$

$$V_0 = 16.560.000$$

Tenons compte à présent des erreurs d'arrondissement ; nous trouvons les valeurs de r_1 de r , s_1 de s et V par les formules (67) et (66), qui correspondent à $c_0 = 0,353a$. Les résultats de ces calculs sont:

$$r_1 = 4,10 ; \quad s_1 = 22,4 \quad \text{et} \quad V = 16.490.000$$

Comme la valeur $V = 16.000.000$ correspond au coefficient de variation = 5% que nous avons utilisé pour commencer nos calculs, on voit que l'accroissement de la variance, causé par les erreurs d'arrondissement, est:

$$\frac{16.490.000}{16.000.000} \cdot 100 - 100 = 3\%$$

5.3- Sondage à deux degrés avec stratification des PSU (plan IV)

L'expression de la variance est donnée ici par la formule

(14) de la seconde partie:

$$\begin{aligned} \bar{V} &= V_1 + V_2 \\ V_1 &= \sum_i \sum_j \frac{N_{ij}^2 (s_{ij} - 1) \sigma_{ij}^2}{N_{ij} - 1} + \frac{N(\bar{x} - 1) \sigma_x^2}{N - 1} \\ V_2 &= 2V_m \left[\sum_i \sum_j \frac{s_{ij}^2 \sigma_{ij}^2}{N_{ij} - 1} - \sum_i \sum_j \frac{s_{ij}^2 \sigma_{ij}^2}{N_{ij} - 1} \right] \end{aligned} \quad (68)$$

La fonction de coût que nous adoptons est :

$$C = \frac{1}{2} + \frac{1}{2} \sum_i \sum_j \frac{b}{s_{ij}} \quad (69)$$

Pour calculer r_o , s_{ijo} , c_o , r_1 , s_{ijl} et V , nous devons rendre minimum l'expression $V + \lambda C$. Mais, pour cette forme générale de notre problème, la solution des équations exprimant que $V + \lambda C$ est minimum est si compliquée que nous nous contenterons d'étudier, ci-dessous, deux cas particuliers.

5.3-1. Premier Cas Particulier-

Dans ce paragraphe, nous appliquons le résultat obtenu ci-dessus au plan de sondage, établi par l'auteur pour une enquête sur l'eau utilisée pour l'irrigation dans les villages de l'Iran, dans sa thèse pour le grade de M.Sc. en Statistique du " State College of North Carolina" (voir Ref. 9).

Le plan de ce sondage était le suivant:

1) Le pays de l'Iran sera divisé en N PSU, chacune formée par la réunion de plusieurs districts ruraux (chaque district rural de l'Iran contient, en moyenne 100 villages) ; ce regroupement

des districts est fait de telle sorte que les nombres de villages de toutes les PSU soient autant que possible égaux entre eux afin que la variance entre les totaux des PSU soient aussi petite que possible. Une PSU sur t sera soumise à l'enquête.

2) Les villages de chaque PSU choisie seront stratifiés, d'après les nombres de leurs habitants, en s strates.

Nous avons déjà indiqué, par nos études, que la composante de la variance (entre les quantités d'eau utilisées par les villages) causée par la variation de la grandeur des villages est beaucoup plus grande que celle due à des différences entre régions géographiques. C'est pourquoi nous avons stratifié les villages d'après le nombre de leurs habitants.

3) Dans la strate n° ij , nous soumettons à l'enquête un village sur r_{ij} . Les r_{ij} doivent être proportionnels aux inverses des écarts-types σ_{ij} entre les villages à l'intérieur des strates.

4) Les seules données statistiques que nous avions à notre disposition étaient:

- a) les résultats de " l'enquête TD".
- b) Les nombres de villages de chaque district de l'Iran, et
- c) Les moyennes arithmétiques par village des quantités d'eau utilisées pour irrigation, pour chaque district.

Nous avons compensé le manque de données concernant les districts de l'Iran, en appliquant ici les résultats de "l'enquête TD". Ces résultats sont les suivants:

a) L'écart-type, σ_{ij} , de chaque agglomération de villages est proportionnel à η_{ij} , moyenne arithmétique des quantités d'eau de ces villages, c'est-à-dire:

$$\sigma_{ij} = \beta \eta_{ij} \quad (70)$$

où

$$\beta \approx 0.66$$

b) Nous avons admis que:

$$\eta_{ij} = \tau_j \eta_i \quad (71)$$

(71)

pour tous les districts ou réunions de districts de l'Iran, où τ_j varie d'une strate à l'autre, mais non pas d'une PSU à l'autre.

c) Nous avons admis que

$$N_{ij} = p_j N_i, \quad \sum_j p_j = 1 \quad (72)$$

(72)

pour toutes les PSU, où p_j varie d'une strate à l'autre, et non pas d'une PSU à l'autre ; et nous avons démontré que

$$\sum_j \tau_j p_j = 1$$

ce qui nous a permis de simplifier beaucoup les formules.

d) Nous avons admis qu'on se rapprochait de l'optimum en prenant les fractions de sondage proportionnelles aux écarts-types, comme c'est le cas pour un échantillon stratifié. Et par conséquent nous avons posé:

$$z_{ij} = K/\sigma_{ij}$$

c'est-à-dire (d'après les formules précédentes)

$$z_{ij} = K/\beta \tau_j \eta_i \quad (73)$$

K étant un certain paramètre.

En substituant les formules (70), (71), (72) et (73) dans le système (68) et en confondant $N_i - 1$ et N_i , $N - 1$ et N , nous avons, en première approximation:

$$\begin{cases} V_1 = z K \beta X - z A + N(z-1) \sigma_X^2 \\ V_2 = z K^2 B \\ C = a \left(\frac{1}{z} + \frac{E}{z K} \right) \end{cases} \quad (74)$$

où :

$$A = \beta^2 \sum_i N_i \eta_i^2 \sum_j \rho_j \tau_j^2$$

$$B = \left[\frac{\beta N}{\beta^2} - \sum_i \frac{1}{N_i} \sum_j \frac{1}{\rho_j} \right] V_m$$

$$E = \frac{\beta}{a} \sum_i \sum_j \eta_i \tau_j b = \frac{b\beta}{a} \sum_i \eta_i \sum_j \tau_j$$

bl) Pour calculer r_o , k_o et c_o , nous rendons minimum l'expression $V_1 + \lambda c$. Les résultats de ce calcul sont:

$$K_0 = \sqrt{\frac{E(N\sigma_X^2 - A)}{\beta X}} \quad (75)$$

$$z_0 = \frac{V_1 + N\sigma_X^2}{K_0 \beta X - A + N\sigma_X^2} \quad (76)$$

et

$$C_0 = a \left(\frac{1}{z_0} + \frac{E}{z_0 K_0} \right) \quad (78)$$

$$b_2) V'_0 = V_0 + r_0 K_0^2 B$$

b₃) Afin de calculer r_1 , k_1 et V , nous rendons minimum l'expression $V + \lambda c$, ce qui donne:

$$2BK_1^3 + (\beta X + EB)K_1^2 - (N\sigma_X^2 - A)E = 0 \quad (79)$$

$$z_1 = \frac{a}{c_0} \left(1 + \frac{E}{K_1} \right) \quad (80)$$

et

$$V = z_1 K_1 \beta X + z_1 A + z_1 K_1^2 B + N(z_1 - 1) \sigma_X^2 \quad (81)$$

On voit que les formules (75) à (81) ont les mêmes propriétés mathématiques que celles du plan II.

Exemple Numérique

pour le premier cas particulier

Pour appliquer le plan de ce paragraphe, nous utilisons les mêmes données que pour la thèse citée plus haut, lesquelles sont:

$$\begin{aligned}
 N &= 76, & N_1 &= 42.000, & s &= 4, & \sum_j p_j t_j &= 1,347 \\
 \sum_i N_i \eta_i^2 &= 185.000.000, & &= 0,18, & &= 111, \\
 b/a &\stackrel{v}{=} 30, & \sum_i \eta_i &= 5.000, & \sum_j t_j &= 8,3; & X &= 2,683.000 \\
 E(N_1^2) &= 308.000, & \sigma_{\eta} &= 150, & E(\eta_1^2) &= 4.600, \\
 \sigma_{N_1}^2 &= 8.000 & \beta &= 0,66.
 \end{aligned}$$

En utilisant les données ci-dessus et les formules (74) à (81), nous avons trouvé les résultats suivants:

$$\begin{aligned}
 V_0 &(\text{pour un coefficient de variation } 5\%) = 18.000.000.000, \\
 r_0 &= 2,08, & k_0 &= 3.100, & c_0 &= 0,9a, & V_0' &= 19.130.000.000 (\text{donc} \\
 V_0' / V_0 \% &= 106,2\%), & r_1 &= 22,1, & k_1 &= 2745; \\
 V &= 19.070.000.000 (\text{donc } V / V_0 \% = 106,0 \%).
 \end{aligned}$$

On voit que, si dans le calcul de r et k , on tient compte des erreurs d'arrondissement, on ne peut pas avoir de gain pratique de variance.

5.3.2 - Deuxième Cas Particulier:

Le plan de l'enquête ici est:

- 1) La population se divise en N PSU dont $1/r$ seront soumises à l'enquête.
- 2) Chaque PSU sera stratifiée en 3 strates dont la première sera énumérée, et les deuxième et troisième seront soumises à

l'échantillonnage, respectivement avec les fractions de sondage $1/s$ et $1/t$.

Avec ce plan et les hypothèses $N_{ij} \neq 1$, $N_{ij} \neq N$, $\sigma_{i2}^2 = \sigma^2$ et $\sigma_{i3}^2 = \beta \sigma^2$, les formules (68) et (69) seront en première approximation remplacées par:

$$\begin{cases} V_1 = z(\beta-1)A + z(t-1)B + N(z-1)\sigma_x^2 \\ V_2 = z\beta E + ztF \\ = \frac{a}{z} + \frac{b}{zt} + \frac{d}{zt} \end{cases} \quad (82)$$

ou

$$A = N_{o2} \sigma^2, \quad B = N_{o3} \beta \sigma^2, \quad N_{oJ} = \sum_i N_{ij}$$

$$E = V_m \left[\sum_i n_{i2}^2 - \sigma^2 \sum_i \frac{1}{N_{i2}} \right]$$

$$F = V_m \left[\sum_i n_{i3}^2 - \beta \sigma^2 \sum_i \frac{1}{N_{i3}} \right]$$

b₁) Afin de calculer les valeurs r_o , s_o , t_o et C_o , pour la valeur désirée de la variance V_o , nous rendons minimum l'expression $V_1 < C$, ce qui donne:

$$\frac{b}{a} B t_o^2 - 2 \left[(A+B-\sigma_x^2) \frac{b}{a} + A \beta \right] t_o - \frac{d}{a} A \beta_o^2 = 0 \quad (83)$$

$$\frac{d}{a} A \beta_o^2 - 2 \left[(A+B-\sigma_x^2) \frac{d}{a} + B t_o \right] \beta_o - \frac{b}{a} B t_o^2 = 0 \quad (84)$$

$$t_0 = \frac{V_0 + N E_x}{(s_0 - 1)A + (t_0 - 1)B + N E_x} \quad (85)$$

et

$$c_0 = \frac{1}{t_0} \left(a + \frac{b}{s_0} + \frac{d}{t_0} \right) \quad (86)$$

Les équations (83) et (84) sont indépendantes de V_0 , valeur désirée de V . L'équation (83) est du deuxième degré en t_0 et son déterminant est toujours positif. Pour chaque valeur de s_0 on peut donc trouver pour t deux valeurs dont nous choisissons la plus grande. De cette façon, nous pouvons tracer la courbe de variation de t_0 par rapport à s_0 d'après l'équation (83). De la même façon nous pouvons tracer la courbe de variation de s_0 par rapport à celle de t_0 d'après l'équation (84). Les coordonnées du point d'intersection de ces deux courbes donnent les valeurs optimum de s_0 et t_0 .

b₂)

$$V_0' = V_0 + t_0^2 s_0^2 E + t_0^2 F$$

b₃) Afin de calculer les valeurs r_1 , s_1 , t_1 et V , pour la valeur fixe c_0 du coût, nous rendons minimum l'expression

$V + \lambda c$, ce qui donne:

$$\begin{aligned} & \frac{b}{a} B t_1^2 - \frac{d}{a} (2E s_1 + A) s_1^2 \\ & - 2 \left[(A+B - \sigma_x^2) \frac{b}{a} \right. \\ & \left. + (2E s_1 + E \frac{b}{a} + A) s_1^2 \right] t_1 = 0 \end{aligned} \quad (87)$$

$$\begin{aligned} & \frac{d}{a} A s_1^2 - \frac{b}{a} (2F t_1 + B) t_1^2 \\ & - 2 \left[(A+B - \sigma_x^2) \frac{d}{a} \right. \\ & \left. + (2F t_1 + F \frac{d}{a} + B) t_1^2 \right] s_1 = 0 \end{aligned} \quad (88)$$

que nous pouvons résoudre de la façon déjà dite pour les équations (83) et (84). Nous aurons:

$$z_1 = \frac{1}{c_0} \left(a + \frac{b}{s_1} + \frac{d}{t_1} \right)$$

$$V = z_1 (s_1 - 1) A + z_1 (t_1 - 1) B + N (z_1 - 1) \sigma_x^2 + z_1^2 \frac{s_1}{1} E + z_1^2 \frac{t_1}{1} F \quad (89)$$

- Deuxième Exemple Numérique -

1) Nous appliquons ce plan à l'enquête qui a pour but l'estimation du nombre total des ovidés (moutons et chèvres) de l'Iran. Les données qui seront utilisées pour ce plan proviennent de deux sources:

a) Celles concernant le nombre des ovides, proviennent de l'enquête par énumération de la province de Fars en 1952 (voir Ref.8). Nous avons supposé (faute de mieux) que les paramètres

calculés pour la région de Fars sont utilisables pour tout le pays.

b) celles concernant le nombre de villages, sont les mêmes que pour l'exemple précédent.

2) Nous avons utilisé les mêmes PSU que pour l'exemple précédent.

3) Nous stratifions chaque PSU en trois strates, d'après les nombres des habitants des villages: (1), 851 et plus, (2) 1 à 250 et (3) 251 à 850. Nous allons soumettre à l'enquête tous les villages de (1), un village sur s des (2), et un village sur t des (3).

4) Nous recensons les ovidés de chaque village choisi.

Les données qui ont servi à nos calculs sont:

$$N_{02} = 28.000, N_{03} = 19.000, \sigma^2 = 22.000, \beta = 4,7,$$

$$N = 76, E(N_1^2) = 308.000, \sigma_{\xi_{10}}^2 = 4.900, E(\xi_{10}^2) = 100.000$$

$$\sigma_{N_1}^2 = 8.000, \sum \xi_{10}^2 = 2.000.000, \sum \xi_{30}^2 = 22.000.000$$

$$\sum \frac{1}{N_{12}} = 0,44, \sum \frac{1}{N_{13}} = 0,36, b/a = 10, d/a = 30 \text{ et}$$

$$X = 15.000.000$$

En portant ces données dans les formules (82) à (89), nous obtenons:

$$V_0 \text{ (pour un coefficient de variation 5\%)} = 562.000.000.000,$$

$$s_0 = 59,5, t_0 = 55,3, r_0 = 2,16, c_0 = 0,795 a,$$

$$V_0' = 575.700.000.000 \text{ (donc } V_0' / V_0 \% = 102\%), s_1 = 58, t_1 = 54$$

$$r_1 = 2,17 \text{ et } V = 571.000.000.000 \text{ (donc } V/V_0 \% = 101,4\%).$$

On voit qu'ici les erreurs d'arrondissement sont plus petites que dans le cas particulier précédent. La raison en est,

probablement, qu'ici, dans les strates dont les effectifs sont petits (c'est-à-dire les strates n° 1), nous avons énuméré et non pas sondé.

5.4 - 2PSU (en tout), tirages exhaustifs, probabilités inégales
(plan V) -

En utilisant les mêmes approximations qu'au paragraphe 4.3 et en remarquant que:

$$\sum_i \sum_{i \neq i'} X_i X_{i'} = (\sum_i X_i)^2 - \sum_i X_i^2$$

nous transformons la formule (18) de la seconde partie en :

$$V(\hat{X}) = V_1 + V_2 \quad (90)$$

avec

$$V_1 = NN_0 \sigma^2 \left(1 - \frac{1}{N}\right) + \frac{N_0}{2} \sum_i N_i \bar{X}_i^2 - \frac{1}{2} \sum_i X_i^2 - \frac{1}{2} X^2$$

$$V_2 = A^2 N_0 \sigma^2$$

$$A = \frac{2 \bar{V}_m}{N_0} \left[\sum_i N_i \bar{X}_i^2 - N_0 \sigma^2 \right]$$

b₁) Dans le présent cas, il n'est pas possible de rendre minimum l'expression $V_1 + \lambda c$. Nous rendons donc minimum le produit cV_1 , où la fonction de coût est:

$$c = a + b/r \quad (91)$$

a représente les frais généraux de l'enquête pour deux PSU moyennes et b est le coût de l'enquête pour tous les éléments

de deux PSU moyennes.

Comme résultat de ce calcul, nous trouvons seulement la valeur " optimum indépendante" définie au paragraphe 1.

La valeur " optimum indépendante" r_0 est

$$r_0 = \sqrt{\frac{b}{a} \frac{N_0 \sum N_i \bar{S}_{i0} - \sum X_i^2 - NN_0 \bar{S}^2}{2 \bar{S}^2 N N_0}} \quad (92)$$

et nous avons dit que, si nous voulons trouver des valeurs optimum, c_0 et V_0 sont imposés par le plan et la structure de la population.

$$b_2) \quad V'_0 = V_0 + A r_0^2$$

b3) Calcul de la valeur optimum de r quand nous tenons compte des erreurs d'arrondissement.

Ici nous rendons minimum l'expression $c(V_1 + V_2)$; ce qui conduit, après simplification à une équation du troisième degré en r_1

$$\frac{r_1^2}{1} + \left[\frac{b}{2a} + \frac{N_0 \bar{S}^2}{2NA} \right] \frac{r_1^2}{1} - \frac{b}{a} \frac{1}{4AN_2} (N_0 \sum N_i \bar{S}_{i0} - \sum X_i^2 - \bar{X}^2 - NN_0 \bar{S}^2) = 0 \quad (93)$$

En utilisant les formules (93), (90) et (91), nous pouvons calculer les valeurs optimum r_1 et V qui correspondent au cas où nous tenons compte des erreurs d'arrondissement.

- Application numérique -

Appliquons encore la théorie de ce paragraphe aux données de " l'enquête TD". Ces données sont:

$$\sigma^2 = 0,8, \quad N = 1.281, \quad N_0 = 72.000, \quad \sum N_i \xi_i^2 = 157.000$$

$$\sum X^2 = 12.000.000, \quad X = 80.000 \quad \text{et} \quad b/a = 10.$$

En portant ces données dans les formules (90) à (93), nous avons trouvé:

$$r_0 = 18,2; \quad c_0 = 0,155a, \quad V_0 = 3.875.000.000,$$

$$V'_0 = 4.060.000.000 \quad (\text{donc } V'_0 / V_0 = 105\%).$$

Comme nous l'avons dit, ces valeurs sont des "optimum indépendantes". Le calcul de r_1 et V est donc indépendant de celui de r_0 , V_0 et c_0 .

D'après les formules (93), (90) et (91), nous trouvons:

$$r_1 = 15,8, \quad c_1 = 0,1595a, \quad V = 3.813.000.000$$

Afin de comparer les deux séries de résultats obtenus, calculons $I = C V$ pour chaque série. Nous avons:

$$I_0 = c_0 V_0 = 630.000.000a \quad \text{et} \quad I_1 = c_1 V = 609.000.000a$$

ce qui montre que, au moins théoriquement, nous aurons un certain gain si nous tenons compte des erreurs d'arrondissement quand nous calculons la valeur optimum de r .

5.5 - Stratification puis une PSU de chaque strate, avec probabilités inégales (plan VI) -

En utilisant les mêmes approximations qu'au paragraphe 4.4 et en remarquant que:

$$\sum_K \sum_c \frac{1}{N_K} \left(\frac{X_{Kc}}{N_K} - \bar{X}_K \right)^2 = \sum_K \sum_c \frac{X_{Kc}^2}{N_K} - \sum_K \bar{X}_K^2$$

nous pouvons transformer la formule (20) de la seconde partie en:

$$\begin{cases} \bar{V} = V_1 + V_2 \\ V_1 = \bar{N}_0 N_{00} \sigma^2 e^2 + B \\ V_2 = A e^2 \bar{N}_0 \end{cases} \quad (94)$$

où

$$A = V_m \left[\sum_K \frac{1}{N_{K0}} \sum_K N_{K0} \sum_K N_{K0} \sigma^2 - \sigma^2 \sum_K \frac{N_K}{N_{K0}} \right]$$

et

$$B = \sum_K N_{K0} \sum_K N_{K0} \sum_K N_{K0} \sigma^2 - \sum_K X_K^2 - \sigma^2 \sum_K N_K N_{K0}$$

b₁) Ici aussi, comme pour le plan précédent, nous rendons minimum le produit cv_1 , où

$$c = a + b/r \quad (95)$$

a représente les frais généraux de l'enquête pour N PSU, et b est le coût de l'enquête pour tous les éléments de N PSU.

Le résultat de ce calcul est:

$$e_0 = \sqrt{\frac{b}{a} \frac{1}{\bar{N}_0 N_{00} \sigma^2}} \quad (96)$$

De même que pour le plan précédent, r_0 est ici un "optimum indépendant".

$$b_2) V'_0 = V_0 + AN^2 \cdot r_0^2$$

$b_3)$ Cette fois, nous rendons minimum le produit $c(V_1 + V_2)$ ce qui donne l'équation de troisième degré en r :

$$r_1^3 + \left[\frac{b}{a} + \frac{N_{00} \sigma^2}{2AN_0} \right] r_1 - \frac{bB}{2aAN_0^2} = 0 \quad (97)$$

En utilisant les formules (97), (94) et (95), nous calculons les valeurs optimum r_1 et V correspondant au cas où nous tenons compte des erreurs d'arrondissement.

Application numérique

Reprenons les données de "l'enquête stratifiée TD" qui sont:

$$\begin{aligned} \sigma^2 &= 0,8, N_{..} = 72.000, \sum_K N_K N_{K_0} = 18.000.000, \\ \sum_K N_K \sum_{K_0} N_{K_0}^2 - \sum_K X_K^2 &= 330.000.000, \sum_K \frac{1}{N_K} \sum_{K_0} N_{K_0}^2 = 11, \\ \sum_K N_K / N_{K_0} &= 0,09 \text{ et } b/a = 10, \text{ ce qui donne} \\ A &= 0,92 \text{ et } B = 316.000.000 \end{aligned}$$

Les formules (96), (94) et (95) donnent:

$$r_0 = 14,6, c_0 = 1,887a, V_0 = 532.000.000 \text{ et}$$

$$V'_0 = 544.700.000 \text{ (donc } V'_0 / V_0 \% = 104\%).$$

Les formules (97), (94), (95) donnent:

$$r_1 = 13,6; c_1 = 1,731a; V = 528.000.000$$

De même qu'au paragraphe précédent, ces deux séries de résultats sont indépendantes l'une de l'autre et du choix que nous

avons fait de V_0 ou c_0 . Pour les comparer nous calculons:

$$I_0 = c_0 V_0 = 914.000.000a \text{ et } I_1 = c_1 V = 910.000.000a$$

Concluons: le gain est insignifiant.

5.6 - Sondage à deux degrés avec probabilités inégales et tirages bernoulliens au premier degré (plan VII)-

En utilisant les approximations adoptées au paragraphe 4.5 et considérant que :

$$\sum h_i \left(\frac{X_i}{h_i} - \bar{X} \right)^2 = \sum \frac{X_i^2}{h_i} - \bar{X}^2$$

nous transformons la formule (23) de la seconde partie en :

$$V = V_1 + V_2 \quad (98)$$

$$V_1 = \frac{1}{n} \sigma^2 N_0 (2 - N) + \frac{N_0}{n} \sum N_i \bar{X}_{i0}^2 - \frac{\bar{X}^2}{n}$$

$$V_2 = \frac{V_m \bar{X}^2}{n^2} + \frac{V_m \bar{X}^2}{N_0 n} \sum N_i \bar{X}_{i0}^2 + \frac{2 V_m}{N_0^2} \sum X_i^2$$

b₁) La fonction de coût ici est:

$$c = an + bnN/r \quad (99)$$

où a représente les frais généraux de l'enquête pour une PSU moyenne, et b est le coût de l'enquête pour tous les éléments d'une PSU moyenne.

b₂) Afin de calculer n_0 et r_0 , nous rendons minimum l'expression $V_1 + \lambda c$; ce qui conduit à:

$$z_0 = \sqrt{\frac{Nb}{a} \left[\frac{\sum N_i \xi_{i0}^2}{\sigma^2} - \frac{\bar{X}^2}{N_0 \sigma^2} - N \right]} \quad (100)$$

$$n_0 = \frac{\sigma^2 N_0 (z_0 - N) + N_0 \sum N_i \xi_{i0} - \bar{X}}{V_0} \quad (101)$$

$$c_0 = a n_0 + \frac{b N n_0}{z_0} \quad (102)$$

b₂)

$$V'_0 = V_0 + V_m \left[\frac{\bar{X}^2}{n_0^2} + \frac{z_0^2}{n_0} \sum N_i \xi_{i0}^2 + \frac{z_0^2}{N_0^2} \sum \bar{X}_i^2 \right] \quad (103)$$

b₃) Afin de calculer n₁ et r₁, nous rendons minimum l'expression (V₁ + V₂) + λc, ce qui donne l'équation du troisième degré suivante:

$$\begin{aligned} & \frac{3}{z_1} \left[\frac{2 n_1 V_m}{N_0} \sum N_i \xi_{i0}^2 + 2 \frac{n_1^2}{N_0^2} V_m \sum \bar{X}_i^2 \right] \\ & + \frac{2}{z_1} \left[n_1 N_0 \sigma^2 + n_1 \frac{b}{a} \frac{N}{N_0} V_m \sum N_i \xi_{i0}^3 + 2 n_1^2 \frac{b}{a} \frac{N}{N_0^2} V_m \sum \bar{X}_i^2 \right] \\ & - \left[n_1 \frac{b}{a} N N_0 \sum N_i \xi_{i0}^2 + 2 \frac{b}{a} N V_m \bar{X}^2 \right. \\ & \left. - n_1 \frac{b}{a} N N_0 \sigma^2 - n_1 \frac{b}{a} N \bar{X}^2 \right] = 0 \quad (104) \end{aligned}$$

Pour utiliser l'équation (104), 1) nous donnons plusieurs valeurs convenables à n₁ et nous calculons les valeurs correspondantes de r₁ et celles de c par la formule (99).

2) nous trouvons, par interpolation, les valeurs de n_1 et r_1 qui correspondent à c_0 .

- Application numérique -

Les données de "l'enquête TD" que nous avons utilisées pour ce plan sont:

$$\sum X_i^2 = 12.000.000, \quad \sum N_i \xi_i^2 = 157.000, \quad X = 80.000, \\ N = 1281, \quad N. = 72.000, \quad b/a = 10 \quad \text{et} \quad \sigma^2 = 0,08.$$

Les résultats des calculs sont:

$$V_0 \text{ (pour un coefficient de variation 5\%)} = 16.000.000$$

$$r_0 = 33.000, \quad n_0 = 462, \quad c_0 = 592 \text{ a,}$$

$$V'_0 = 16.670.000 \text{ (donc } V'_0 / V_0 = 104,0\%)$$

$$r_1 = 32.200, \quad n_1 = 425,0 \text{ et}$$

$$V = 16.600.000 \text{ (donc } V/V_0 = 103,6\%).$$

Le gain est insignifiant.

6 - Resumé de la 3ième partie.

6.1-Résultats généraux des problèmes du type 1 (répartition optimum)

Afin d'étudier l'influence des erreurs d'arrondissement, dans les cas où l'on recherche une répartition optimum, nous avons choisi les formules (23) et (36) dont la première illustre le cas de la stratification et la deuxième illustre, d'une façon générale, le cas du sondage à plusieurs degrés.

En posant $M_a/M_h = 2$ (voir § 3.1), la formule (23) s'écrit

$$d = \frac{2V_m N^2}{n^2}$$

ce qui nous montre que, lorsqu'il y a stratification, la partie de la variance causée par les erreurs d'arrondissement croît avec N , nombre des strates, et décroît quand l'effectif total n de l'échantillon augmente.

En supposant $\bar{n}_\alpha = \bar{n}$, pour tout α , la formule (36) s'écrit.

$$d = \frac{(k+1)V_m}{2\bar{n}^2}$$

ce qui montre que la variance des sondages à plusieurs degrés croît avec k (nombre des degrés de sondage) et décroît avec l'effectif $\bar{n}$ de l'échantillon pour chaque degré de sondage.

On voit que ces résultats ne sont pas autre chose que ceux que nous avons déjà obtenus à la fin de la deuxième partie.

6.2 - Résultats des problèmes du type 2.

La comparaison mathématique des résultats des problèmes du type 2 n'est pas possible pour nous, même sous la forme abstraite et simplifiée utilisée au § 9 de la 2^{ème} partie.

Nous nous contenterons donc d'utiliser les résultats des calculs numériques. Nous portons les valeurs des $V'_0/V_0\%$ et $V/V_0\%$ que nous avons obtenues pour chaque plan, dans le tableau ci-dessous.

Les définitions de V_0 , V'_0 et V ont été déjà données au § 1, mais répétons-les brièvement:

V_0 est la valeur désirée de la variance, sauf pour les plans V et VI où V_0 est imposée par le plan et la structure de la population (voir " optimum indépendant" § 1).

V'_0 est la variance qu'on obtient si on n'a pas tenu compte des erreurs d'arrondissement dans les calculs des valeurs optimum des fractions de sondage.

$V(=V_1 + V_2)$ est la variance totale quand on a tenu compte des erreurs d'arrondissement dans les calculs des valeurs optimum des fractions de sondage.

	$V'_0/V_0\%$	$V/V_0\%$
Plan II	105,0	104,4
Plan III	103,7	103,0
Plan IV(1° cas particulier)	106,2	106,0
Plan IV(2° cas particulier)	102,0	101,4
Plan V	105,0	101,2
Plan VI	104,0	101,4
Plan VII	104,0	103,6

Pour les plans V et VI, nous avons ajusté les V obtenues au § 5.4 et 5.5 sur le même coût que les V'_0 c'est-à-dire nous avons remplacé V par $V \cdot c_1/c_0$ (voir § 5.4 et 5.5)

Le tableau ci-dessus montre que :

1) Quand on calcule les valeurs optimum des fractions de sondage sans tenir compte des erreurs d'arrondissement, la variance sera (disons) de 2 à 6% plus grande que la valeur désirée V_0 qu'on a portée dans les calculs.

2) Quand on tient compte des erreurs d'arrondissement dans les calculs des valeurs optimum des fractions de sondage ,

théoriquement on obtient un certain gain de variance, mais pratiquement ce gain est trop faible pour annuler la contribution des erreurs d'arrondissement à la variance.

Donc :

3) Pour être sûr que la variance définitive V ne dépassera pas la valeur désirée V_0 , on devrait remplacer, dans les calculs, V_0 par une valeur $V_{01} < V_0$, telle que V'_0 soit justement égale à V_0 .

4^{ème} Partie - Problème de recherches de structure optimum.

1 - Enoncé de la forme générale du problème.

Dans cette partie sont repris les seuls problèmes du type 2 de la troisième partie. Nous verrons, à la fin de ce paragraphe, que les problèmes que nous résoudrons au cours de cette 4^{ème} partie, sont des formes plus générales des problèmes déjà traités.

La première fois qu'on veut procéder à l'enquête par sondage dans une population, on constate que cette population a une structure d'une nature plus ou moins naturelle, c'est-à-dire elle contient des unités de différents ordres de sondage, qui sont indiquées par la nature ou par l'ordre social : un village, un ménage, une usine, etc... Généralement, ce sont ces unités qu'on adopte, au moins pour la première enquête par sondage. Mais après avoir exécuté une fois le sondage, et même quelque-fois avant le premier sondage, on constate que ces unités naturelles ne sont pas les meilleures, soit au point de vue variance, soit au point de vue coût, ou les deux.

On est donc conduit à fusionner des unités de même ordre pour former des unités nouvelles plus grandes, et quelquefois (si cela est possible), à partager quelques-unes des unités. Cette partie de la théorie de l'enquête par sondage est, en l'état actuel, la partie la moins développée.

Nous avons déjà indiqué, dans l'introduction, une méthode de calcul de la taille optimum des strates de la ville de Téhéran, telles que la somme des deux termes de la variance,

$V_1 + V_2$, soit minimum (voir § 5 de l'introduction).

Au cours de cette 4^{ème} partie, nous allons étudier ce même problème pour tous les plans de sondage mentionnés dans la deuxième partie.

Nous allons utiliser le procédé déjà employé dans notre introduction, pour chercher la taille optimum des strates, - c'est-à-dire que, pour les unités de chaque ordre qui doivent prendre la taille optimum, nous définissons un nombre q de la façon suivante:

On réunit plusieurs unités naturelles, de sorte que (1) chaque unité nouvelle renferme, en moyenne, q unités naturelles, et (2) les nombres d'éléments que renferment les unités nouvelles soient autant que possible égaux entre eux.

Nous avons posé la condition (2), non à cause des erreurs d'arrondissement, mais pour rendre les autres erreurs autant que possible plus petites; et nous ferions ainsi, même si nous ne voulions pas tenir compte des erreurs d'arrondissement. Nous posons la condition (2) quand les unités que nous voulons réunir sont des unités de sondage, mais non lorsqu'elles sont des strates.

Pour chaque plan de sondage nous allons suivre des étapes suivantes de calcul.

a) Nous cherchons comment introduire q dans les différents paramètres de la population figurant dans les expressions de c et de $V = V_1 + V_2$. C'est la partie la plus difficile de nos calculs; car les recherches que les auterus ont déjà faites sur

le problème de la tialle des unités de sondage ne sont pas assez complètes, ni leurs formules correspondantes assez simples, pour qu'elles puissent nous permettre de transformer nos formules, en y introduisant q de sorte que les calculs d'optimum soient pratiquement possibles. Nous sommes donc obligés d'adopter des méthodes approchées, mais simples (autant que possible), qui nous donnent seulement une idée générale des solutions de nos problèmes.

b) Choisir les limites de variation de q . A cause de la nature plus ou moins approchée de nos formules quand nous y introduisons q , celles-ci seront pratiquement exactes si nous les utilisons pour étudier de petites variations de q . D'autre part, la nature et le nombre total des unités de sondage sont tels que nous ne sommes pas libres d'adopter n'importe quelle valeur de q . Pour ces deux raisons, nous sommes toujours obligé de choisir entre quelles limites on peut faire varier q . Cependant dans les calculs numériques, nous laissons q varier dans des intervalles plus grands, car (1), nous ne sommes pas capables d'indiquer les limites de variation exactes, et (2) on pourra mieux voir ainsi les effets de la variation de q et ses tendances, tandis qu'on reste toujours libre de ne retenir qu'une partie de l'intervalle étudié.

c) Nous choisissons quelques valeurs convenables de q dans l'intervalle de variation; et pour chacune, nous calculons V_{q1} , V_{q2} et $V_q = V_{q1} + V_{q2}$, en rendant minimum V pour la valeur c_0 de c déjà calculée au bl, §4, de la 3ème partie. Notons que

V_1 , V_2 et $V = V_1 + V_2$ que nous avons calculées au b3 du § 4 de la troisième partie, sont respectivement des cas particuliers ($q = 1$) de V_{q1} , V_{q2} et V_q de cette partie.

Nous suivrons ci-après les trois étapes a, b et c pour les plans de sondages étudiés; et nous donnerons seulement les explications complémentaires.

Au cours de cette partie, nous étudierons, au point de vue de la structure optimum, les 6 plans de sondage (II à VII) étudiés dans nos 2ème et 3ème parties.

Nous avons choisi 3 d'entre eux (les plans II, III et VI) pour les applications numériques, en y tenant compte des deux fonctions de coût (18) et (19) introduites au § 2 de la troisième partie.

Pour le plan II nous construisons des PSU nouvelles; pour le plan VI nous construisons des strates nouvelles; tandis que, pour le plan III, nous construisons en même temps des PSU et des strates nouvelles.

2 - Plan II - Deux degrés de sondage avec probabilités égales et tirages exhaustifs.

D'après les formules du § 3 de la seconde partie, la variance de ce plan est:

$$\begin{cases} V = V_1 + V_2 & (1) \\ V_1 = r(s-1)N \cdot s^2 + N(r-1) \sigma_X^2 \\ V_2 = \Lambda r s^2, \text{ ou } \Lambda = \left(\sum_i \bar{y}_i^2 - \sigma^2 \sum_i 1/N_i \right) / 2 \end{cases}$$

a) Ici la solution du problème n'est pas aussi simple que celle de l'introduction, car le terme σ_x^2 qui se présente dans V_q varie avec q , et nous ne connaissons pas la relation entre q et σ_x^2 . Tandis que ce terme n'apparaissait pas dans le calcul de l'introduction. Nous allons donc résoudre cette partie du problème d'une façon approchée, si nous n'avons pas à notre disposition des valeurs estimées assez bonnes de σ_x^2 pour les différentes valeurs de q . Quant aux autres paramètres se trouvant dans la formule (1), nous les représenterons plus ou moins approximativement par des fonctions assez simples de q . Nous allons supposer que les N_i sont les nombres des éléments se trouvant dans les PSU actuelles et que nous appelons les PSU naturelles.

La réunion des PSU sera faite de façon que le nombre des nouvelles PSU soit égal à N/q . Dans ce cas, les paramètres se trouvant dans la formule (1) subiront les changements suivants:

a1) N . (nombre total des éléments de la population) ne change pas.

a2) Pour la variation de σ^2 , nous devrions prendre plutôt la formule $A q g$ déjà utilisée dans l'introduction (et où A ne doit pas être confondu avec A des formules (1)); mais comme ici nous avons besoin de simplifier, nous admettons que σ^2 soit changé en $q \sigma^2$, c'est-à-dire que g soit égal à 1.

a3) $\sum_i \xi_i^2$ (qui se trouve dans l'expression de A , voir la formule 1) sera changé en $\frac{1}{q} \sum_i \xi_i^2$ et $\sum \frac{1}{N_i}$

(qui se trouve aussi dans A) se change en $\frac{1}{q^2} \sum \frac{1}{N_i}$ $\frac{1}{i}$, car N_i se change, en moyenne, en N_i/q .

On va avoir plus de difficultés pour trouver ce que devient σ_x^2 . Rappelons qu'au § 3 de la 2ème Partie, on a posé

$$X_i = N_i \xi_{i0}, \quad X = \sum N_i \xi_{i0}$$

On peut en déduire.

$$\sigma_x^2 = E(N_i)^2 \sigma_{\xi_{i0}}^2 + E(\xi_{i0})^2 \sigma_{N_i}^2$$

à condition que le terme rectangle soit négligeable, ce qui supposerait les N_i et ξ_{i0} indépendants. Nous admettrons que les moyennes pondérée et non pondérée (par N_i) des ξ_{i0} sont très voisines l'une de l'autre; on a bien alors

$$\frac{1}{N_0} \sum_i N_i \xi_{i0} = \frac{1}{N} \sum_i \xi_{i0} = \xi_{00}$$

d'où

$$\sum_i N_i \xi_{i0} = \sum_i N_i \xi_{00}$$

et par conséquent

$$\sum N_i (\xi_{i0} - \xi_{00}) = \sum (N_i - \bar{N}_0) (\xi_{i0} - \xi_{00}) = \text{covariance}$$

$$(N_i, \xi_{i0}) = 0$$

Ainsi, moyennant une hypothèse qui n'est pas plus mauvaise que les autres, on peut écrire

$$\begin{aligned} \sigma_x^2 &= E(N_i)^2 \sigma_{\xi_{i0}}^2 + E(\xi_{i0})^2 \sigma_{N_i}^2 \\ &= \bar{N}_0^2 \sigma_{\xi_{i0}}^2 + \sigma_{N_i}^2 [\sigma_{\xi_{i0}}^2 + E(\xi_{i0})^2] \text{ ou } \bar{N}_0 = E(N_i) \quad (2) \end{aligned}$$

Dans la formule (2), nous admettons que $\sigma_{\xi_{i0}}^2 + E(\xi_{i0}^2)$ ne dépendent pas de q . En fait, leurs variations sont négligeables à côté de celles de $\bar{N}$. et $\sigma_{N_i}^2$ et de l'approximation faite pour arriver à la formule (12).

Le terme $\bar{N}^2$ se change exactement en $q^2 \bar{N}^2$. La partie la plus difficile à traiter est $\sigma_{N_i}^2$. Si dans la formation des PSU nouvelles, nous tirions au sort q PSU et les considérions comme une PSU, le problème serait plus simple. Mais nous avons posé comme condition (2) que les nombres des éléments des PSU nouvelles (que nous désignerons par N'_i), soient autant que possible égaux entre eux. C'est un point qui rend difficile la réalisation pratique d'une structure optimum. En revanche, il implique que la valeur de $\sigma_{N_i}^2$ tende vers zero, lorsque q croît indéfiniment; ce qui fait tendre vers zero tout le deuxième terme du second membre de la formule (2); par conséquent, celui-ci devient négligeable, quand q augmente.

Pratiquement nous calculons cinq valeurs de $\sigma_{N_i}^2$, et nous interpolons (graphiquement) entre elles pour obtenir toutes les autres valeurs. Sur ces cinq valeurs, deux sont évidentes, à savoir:

$$q = 1, \quad \sigma_{N_i}^2 = \sigma_{N_i}^2$$

$$q = N, \quad \sigma_{N_i}^2 = 0$$

Pour obtenir les trois autres, nous avons mis au point deux méthodes.

a 3 I) La méthode (la plus sûre) consiste, dans le cas où nous avons assez de données, à calculer directement, 3 valeurs de $\sigma_{N_i}^2$ pour 3 valeurs de q , en construisant des PSU choisies par tirage au sort. Cette méthode est généralement inapplicable.

a 3 II) La justification de la deuxième méthode (qui doit être utilisée seulement dans le cas où on ne peut utiliser la première par suite du manque de données) contient deux parties.

a 3 III) Si nous désignons la moyenne arithmétique des N_i par $\bar{N}$, d'abord nous supposons que $q\bar{N}$ est supérieur ou égal au plus grand des N_i , et nous essayons de trouver une formule pour la variance de N_i' .

En construisant les nouvelles PSU, notre but est de faire, autant que possible, N_i' égal à $q\bar{N}$. Afin d'arriver à ce résultat, nous réunissons autant de PSU (par exemple m) qu'il faut pour que la différence $N_i' - q\bar{N}$ soit inférieure au plus petit des N_i , parmi tous ceux des PSU se trouvant au voisinage de ce groupe de m PSU; appelons N_{m+1} sa taille. Alors, si:

$\frac{1}{2}N_{m+1} < q\bar{N}$, nous ajoutons cette $(m + 1)^{ième}$ PSU aux PSU déjà réunies; mais si $\frac{1}{2}N_{m+1} > q\bar{N}$, ce groupe des m PSU est adopté comme nouvelle PSU; et si $\frac{1}{2}N_{m+1} = q\bar{N}$ nous déciderons en tirant à pile ou face.

Désignons par $P(q, x)$ la probabilité pour que la taille de la plus petite PSU, qui se trouve au voisinage de la réunion de m PSU, soit x ; désignons par Θ la fraction que nous devons

multiplier par $N_m + 1$ pour obtenir $N'_i - q\bar{N}$. Il semble logique de supposer que s est indépendant des autres paramètres et a une distribution $\text{rect}(0, 1/2)$. Nous avons donc

$$\sigma_N^2 = E[(eN'_i)^2] = E(e^2)E(N'^2_i) = \frac{1}{12} \sum_i p(q, N'_i) N'^2_i \quad (3)$$

Pour calculer $P(q, N'_i)$ d'une façon aussi simple que possible, nous admettons que (1) les tailles des PSU contigües sont indépendantes les unes des autres, (2) notre population est de forme géographique, c'est-à-dire que nous devons choisir la $(m + 1)$ ième PSU, seulement parmi les PSU contigües aux m PSU déjà réunies (sinon nous pourrions rendre la σ_M^2 , presque égale à zéro, en utilisant les plus petites PSU que nous réserverions pour ce but), (3) à mesure que nous avançons dans la construction des nouvelles PSU, le nombre des PSU proches qui peuvent être retenues comme aptes à être la $(m + 1)$ ième PSU, et leurs probabilités restent constants et égaux à leurs valeurs moyennes, et (4) la réunion des m PSU a à peu près la forme d'un cercle de superficie q (l'unité de superficie est donc $\bar{N}$.) et que toujours les PSU situées immédiatement près d'une moitié de sa circonférence sont à notre disposition (les autres ayant été déjà utilisées); nous allons ainsi négliger le fait que, si les PSU étaient, par hasard, toutes grandes, leur nombre à proximité de la circonférence serait nécessairement réduit. Nous désignerons le nombre V de ces PSU par $V(q)$.

D'après nos hypothèses, nous avons

$$V = \frac{1}{2} \left[\pi \left(\sqrt{\frac{q}{\pi}} + 1 \right)^2 - 9 \right] = \frac{\pi}{2} + \sqrt{\pi q} \neq \sqrt{\pi q}$$

Nous avons négligé : $\pi/2$, car pour les petites valeurs de q ce terme ne semble pas convenir, et pour les grandes valeurs, ce terme est négligeable.

Pour calculer la probabilité $P(q,x)$, ou mieux $P(\sqrt{\quad}, x)$, nous utilisons les travaux de GUMBEL(ref.:10) et de FISHER et TIPPETT(ref.: 11).

Si nous admettons que les N_i ont une distribution continue de forme exponentielle, quelle que soit sa distribution, on peut employer la formule

$$dt = e^y - e^y dy \quad (5)$$

où $y = (x - \bar{x}) \sqrt{\frac{f(\bar{x})}{f(x)}}$, $\bar{x}$ = mode de la distribution, et $f(\bar{x})$ est la valeur de la densité au mode.

En théorie, nous devons utiliser la formule (5) sous la forme

$$\sigma_{N'}^2 = \frac{1}{12} \int_0^\infty x^2 e^y - e^y dy \quad (6)$$

où $e^y - e^y dy$ est la probabilité pour que x , (taille de la plus petite de $\sqrt{\quad}$ PSU qui sont à notre disposition pour choisir la $(m + 1)$ ème PSU) soit comprise entre x et $x + dx$. Comme généralement, les données dont nous disposons pour N_i se présentent sous la forme d'une distribution de fréquence; pour appliquer la formule (6), si nous désignons la longueur de la j ème intervalle par I_j , et le milieu de cet intervalle par N_j , nous avons approximativement,

$$\begin{aligned} \sigma_{N_i}^2 &= \frac{1}{12} \frac{\sum_{j=1}^K N_j^2 I_j^2 V(N_j - \bar{N})^2 f(\bar{N}) - e^{V(N_j - \bar{N})} f(\bar{N})}{\sum_{j=1}^K I_j^2 V(N_j - \bar{N})^2 f(\bar{N}) - e^{V(N_j - \bar{N})} f(\bar{N})} \\ &= \frac{1}{12} \frac{\sum_{j=1}^K N_j^2 I_j^2 P(V, N_j)}{\sum_{j=1}^K I_j^2 P(V, N_j)} \end{aligned}$$

où $\bar{N}$ est le mode, $f(\bar{N})$ la densité au mode, et K le nombre de classes.

Comme les valeurs non négligeable de $P(V, N_j)$ sont concentrées autour du mode de N_j , pour pouvoir utiliser cette formule on doit découper des intervalles très fins autour du mode.

à 3 II 2) Maintenant nous prenons le cas où $q\bar{N} < \text{Max.}(N_i)$. Désignons les $N_i' < q\bar{N}$ par N_{1i}' et les $N_i' > q\bar{N}$ par N_{2i}' . Il est clair que les N_{2i}' sont les mêmes que le N_{2i} , c'est-à-dire les tailles des PSU naturelles. Nous utiliserons donc ces deux notations indifféremment. Nous désignons par $N_{1i}, N_{1i}/q\bar{N} = \sum_{i=1}^q N_{1i}/q\bar{N}$ le nombre des PSU nouvelles qui sont formées par les PSU plus petites que $q\bar{N}$, par N_2 le nombre des PSU nouvelles qui sont les mêmes que les PSU naturelles dont les tailles sont plus grandes que $q\bar{N}$. Il est clair que, dans ce cas:

$$\sigma_{N_1}^2 = \frac{N_1}{N_1 + N_2} E(N_{1i} - q \bar{N}_\cdot)^2 + \frac{N_2}{N_1 + N_2} E(N_{2i} - q \bar{N}_\cdot)^2 \quad (8)$$

Le calcul du deuxième terme du second membre de (8) est facile. En fait, il est égal à

$$\frac{1}{N_2} \sum_{j(q \bar{N}_\cdot)}^k F(N_{2j}) (N_{2j} - q \bar{N}_\cdot)^2$$

où $j(q \bar{N}_\cdot)$ = le numéro de l'intervalle dont la limite inférieure est égale à $q \bar{N}_\cdot$.

Mais le calcul de $E(N_{1i} - q \bar{N}_\cdot)^2$ est difficile, car d'abord on pense qu'on doit la poser égale à

$$\frac{1}{N_1} \frac{\sum_{j=1}^{j(q \bar{N}_\cdot)-1} N_j^2 I_j P(V, N_j)}{j(q \bar{N}_\cdot)-1} - \frac{\sum_{j=1}^{j(q \bar{N}_\cdot)-1} I_j P(V, N_j)}{j(q \bar{N}_\cdot)-1}$$

Mais la formule (5) s'y oppose. Pour résoudre ce problème, au moins d'une façon approchée, nous considérerons que $E(N_{2i} - q \bar{N}_\cdot)^2$, au moins pour les petites valeurs de V , est beaucoup plus grande que $E(N_{1i} - q \bar{N}_\cdot)^2$. Donc notre erreur sera petite si nous prenons les $P(V, N_i)$ égales entre elles pour toutes les valeurs de N_{1i} , et écrivons

$$E(N'_{1i} - q \bar{N}_i)^2 = \frac{1}{12 q N_1} \sum_{j=1}^{j(q \bar{N}_i)-1} F_{N_{1j}}^2 \quad (\text{où } q N_1 = \sum_{j=1}^{j(q \bar{N}_i)-1} F_{N_{1j}})$$

Ici, $\frac{1}{q N_1}$ remplace $P(V, N_i)$. Nous avons donc

$$\sigma_{N_i}^2 = \frac{1}{N_1 + N_2} \left[\sum_{j=1}^K \frac{F_{(2j)}(N_{2j} - q \bar{N}_i)^2}{j(q \bar{N}_i)} + \frac{1}{12 q} \sum_{j=1}^{j(\bar{N}_i)-1} \frac{F_{(N_{2j})}^2}{N_{2j}} \right] \quad (9)$$

si $q \bar{N}_i < \text{Max}(N_i)$

Pour utiliser les formules (7) et (9), nous calculons un des trois points par (9), le deuxième pour $q \bar{N}_i = \text{Max}(N_i)$ et le troisième pour $q \bar{N}_i > \text{Max}(N_i)$ par (7). Nous allons illustrer cette méthode au cours de notre calcul numérique.

a3) III. D'après le § 2 de la troisième partie nous allons admettre que la fonction de coût reste la même.

Tout ceci concerne le cas où nous utilisons la formule (18) de la troisième partie. Quand nous voulons utiliser la formule (19), nous devons diviser simplement chaque a par q , dans les résultats de nos calculs.

b) cette étape de calcul sera faite à propos de chaque calcul numérique.

c) Maintenant nous sommes en mesure d'introduire q dans nos formules. Pour les formules (1) de cette partie et (62) et (63) de la troisième partie, nous modifions simplement chaque paramètre, d'après les explications données ci-dessus.

Nous allons retrouver la formule (15) pour une valeur fixe que nous avons calculée au cours de l'étape (b1) du § 3 de la

deuxième partie, et puis nous introduisons q .

Après avoir construit les formules donnant les valeurs optimum de s et r pour des valeurs fixes de q , nous trouverons les valeurs optimum de r et s , pour différentes valeurs de q , et chaque fois, nous calculerons la valeur de V correspondante. La valeur de q qui correspond à la plus petite des valeurs de V ainsi calculées est la valeur optimum de q .

Les formules renfermant q sont

$$\begin{aligned} V_q &= V_{q_1} + V_{q_2} \\ V_{q_1} &= z_q(z_q - 1)qN_0\sigma^2 + \frac{N}{q}(z_q - 1)\left[(qN_0)\sigma_{\bar{s}_0}^2 + \sigma_{N_0}^2(\sigma_{\bar{s}_0}^2 + E(\bar{s}_0^2))\right] \\ V_{q_2} &= A z_q s_q / q \end{aligned} \quad \begin{aligned} (10) \\ (10) \\ (11) \end{aligned}$$

$$\begin{aligned} \frac{2A}{q} s_q^3 + (qN_0\sigma^2 + \frac{b}{q} A) s_q^2 - \frac{b}{q} \left[qN_0\sigma_{\bar{s}_0}^2 + \sigma_{N_0}^2(\sigma_{\bar{s}_0}^2 + E(\bar{s}_0^2)) \right] - qN_0\sigma^2 &= 0 \\ z_q &= \frac{1}{c_0} \left(a + \frac{b}{s_q} \right) \end{aligned} \quad \begin{aligned} (11) \quad (12) \\ (12) \end{aligned}$$

Afin de trouver la valeur de V , quand nous tenons compte des erreurs d'arrondissement, nous substituons les valeurs s_q et r_q dans la formule

$$V_q = z_q(z_q - 1)qN_0\sigma^2 + (z_q - 1)\left[qN_0\sigma_{\bar{s}_0}^2 + \frac{1}{q}\sigma_{N_0}^2(\sigma_{\bar{s}_0}^2 + E(\bar{s}_0^2))\right] + \frac{A}{q} z_q^3 \quad (13)$$

On voit que les deux premiers termes de (13) croissent avec q . Mais son troisième décroît quand q augmente, car s_q^2 diminue, et l'augmentation de r_q est beaucoup moins rapide

que celle de q . Il y a donc une valeur de q qui rend V minimum.

Exemple numérique : Pour appliquer la théorie de ce §, nous utilisons les résultats de "l'enquête TD" déjà utilisés au §4.1 de la 3ème partie, pour calculer r_0 , s_0 et c_0 .

Nous changeons la structure de la population de la façon suivante:

1) Chaque PSU nouvelle sera constituée par plusieurs villages que nous considérons comme PSU naturelles.

2) Dans chaque PSU nouvelle (dont nous avons tiré au sort N/r), nous choisirons, par tirage au sort, μ_i des paysans, où μ_i est obtenu par la méthode A d'arrondissement.

3) A chaque paysan ainsi choisi, on demandera la superficie des terres où il a cultivé du blé et de l'orge.

Notre problème ici est de trouver les valeurs optimum de q et s . Les données numériques sont:

$$N = 1.281, N_0 = 72.000, \sigma^2 = 0,8, \sigma_{\xi_i}^2 = 1$$

$$E(\xi_i^2) = 2,21, \sum \xi_i^2 = 2.850, \sigma_{N_i}^2 = 1.200,$$

$$\sum \frac{1}{N_i} = 23, b/a = 10.$$

Nous avons trouvé, pour un coefficient de variation égal à 5% ($V = 16.000.000$), $C_0 = 0/56$ a.

Pour ces calculs, nous devons d'abord avoir les valeurs de σ_N^2 pour les valeurs choisies de q : Nous avons calculé ces valeurs de la façon suivante:

Nous avons utilisé la répartition suivante des villages des deux provinces suivant le nombre de paysans par village :

Nombre de paysans N_i	Milieu de l'intervalle N_j	Nombre de villages
1 - 20	10	150
21 - 60	40	616
61 - 100	80	449
101 - 160	130	51
161 - 240	200	12
241 - 360	300	3

$$\begin{aligned} \check{N} &= 42 \\ &(\text{valeur} \\ &\text{modale de} \\ &N_i) \\ f(\check{N}) &= 0,128 \end{aligned}$$

Nous avons calculé $\sigma_{N_i}^2$ pour 3 valeurs suivantes de V :

$V = 2.31$ correspondant à la valeur $q = 1,33$ en prenant

$q\check{N} = 100$, limite supérieure de l'intervalle 61-100. Nous avons trouvé $\sigma_{N_i}^2 = 540$ d'après la formule (9).

$V = 4.5$ correspondant à $q = 6,4$ en prenant $q\check{N} = 360$. Nous avons trouvé $\sigma_{N_i}^2 = 150$ d'après la formule (7).

$V = 10$, correspondant à $q = 32$. Nous avons trouvé $\sigma_{N_i}^2 = 140$ d'après la formule (7).

Pour utiliser la formule (7), nous avons divisé la classe 21-60 qui contient le mode, en 7 classes plus fines.

Nous avons porté ces trois points, ainsi que les deux points ($q = 1$, $\sigma_{N_i}^2 = 1.200$) et ($q = 1.281$, $\sigma_{N_i}^2 = 0$), sur un graphique à coordonnées doublement logarithmiques (figure 4).

Sur la figure 4, nous avons relevé les valeurs de $S_N'^2$ pour les valeurs choisies de q ; puis nous avons calculé μ_q , s_q et V_q pour $C_0 = 0,56$ a, à l'aide des formules (11), (12) et (13) ; enfin nous les avons portés dans les deux tableaux suivants:

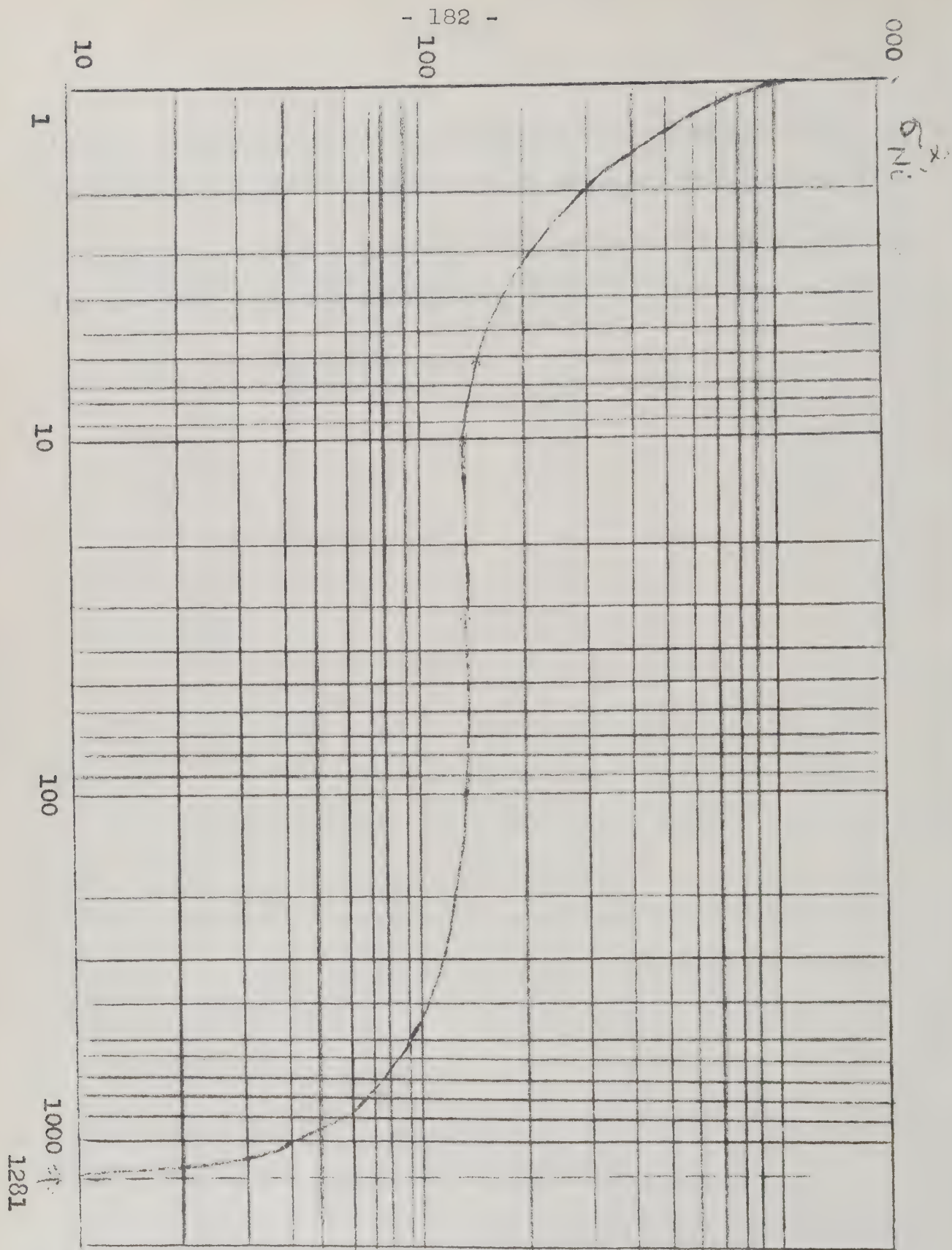


Figure 4 - Graphique de $\sigma_{N_i}^2$

q	$\hat{C} - C = \frac{a}{r} + \frac{b}{rs}$		$\hat{C} - C = \frac{a}{qr} + \frac{b}{rs}$		
	s_q	r_q	s_q	r_q	s_q si $r_q = 1$
1	34,3	2,31	34,3	2,31	-
2	27,2	2,34	38,0	1,50	-
5	26,3	2,46	58,0	0,66	27,8
10	26,3	2,46	83,0	0,39	21,8
20	26,3	2,46	117,0	0,24	18,6

Pour éviter les cas pathologiques ($r_q < 1$), nous avons arrondi les valeurs $r < 1$, et renouvelé le calcul des s_q correspondants. Alors nous avons calculé les V_q , en tenant compte de ces derniers changements, et les avons portés dans le tableau suivant.

q	$C = \frac{a}{r} + \frac{b}{rs}$			$C = \frac{a}{qr} + \frac{b}{rs}$		
	V_{q1} (millions)	V_{q2} (Mille)	$V_q = V_{q1} + V_{q2}$ (Millions)	V_{q1} (Mille)	V_{q2} (mille)	$V_q = V_{q1} + V_{q2}$ (Millions)
1	16,1	645	16,74	16,1	645	16,74
2	20,6	213	20,81	11,0	257	11,26
3	47,2	81	47,28	7,7	27	7,73
4	93,9	41	93,94	11,9	11	11,91
5	189,6	20	189,62	20,2	4	20,20

On voit que, si l'on adopte la fonction de coût $c = a/r + b/rs$, à partir de $q = 5$ les valeurs de s_q et r_q restent les mêmes, tandis qu'avec la fonction de coût $c = a/qr + b/rs$, s_q s'accroît avec q tandis que r_q décroît.

Quant à la variation de V , avec la première fonction de coût, le résultat pour $q = 1$ est meilleur que pour les autres valeurs de q . Mais nous avons fait figurer dans le tableau les variations des deux termes V_{q1} et V_{q2} qui composent V , pour montrer comment un de ces termes augmente rapidement avec q , tandis que l'autre décroît tout aussi rapidement, quand q augmente. Ici le dernier terme est d'ordre beaucoup plus petit que le premier. Mais si l'erreur d'arrondissement était plus grande qu'à présent, le point optimum correspondrait à $q > 1$.

Avec la deuxième fonction de coût, au contraire, tous les deux termes, V_{q1} et V_{q2} , décroissent quand $r > 1$ et q croît.

On voit donc que la fonction de coût que nous adoptons pour nos calculs a une grande influence sur les résultats.

Cependant, on voit que cette différence des résultats se manifeste seulement dans la variation de V_{q1} ; tandis qu'avec l'une ou l'autre des fonctions de coût, le mode de la variation de V_{q2} reste la même: V_{q2} décroît rapidement quand q augmente.

3 - PLAN III - Sondage stratifié à 2 degrés - avec probabilités égales et tirages exhaustifs.

Dans le cas présent nous sommes libres de changer, en même temps la grandeur des PSU et celle des strates, indépendamment

l'une de l'autre à condition que le nombre total des éléments de la population reste constant. Nous pouvons donc introduire deux paramètres q et p dont la définition est analogue à celle donnée au § 1 et 2. Ici q sera le nombre moyen des PSU naturelles et contigües qui doivent former une PSU nouvelle plus grande, et p est le nombre moyen des strates naturelles et contigües qui doivent former une strate nouvelle plus grande.

a) En adoptant les approximations admises au § 2, nous pouvons dire qu'après avoir introduit p et q , les paramètres de la formule (69) de la 3ème partie subiront les changements suivants:

1) N sera changé en $N./q_2$

2) σ^2 sera changé en $q \frac{\sigma^2}{2}$

3) $\frac{\sigma^2}{x} = \bar{N}^2 \dots \sigma_{Kio}^2 + \sigma_{N_{Kio}}^2 \left[\bar{\xi}_{co}^2 + 2 \sigma_{\xi_{Kio}}^2 \right]$
sera changé en $p q \bar{N}^2 \dots \sigma_{Kio}^2 + \sigma_{N_{Kio}}^2 \left[\bar{\xi}_{co}^2 + 2 \sigma_{\xi_{Kio}}^2 \right]$

où $\bar{N} \dots$ est la moyenne arithmétique des N_{ki} dans toute la population, $\bar{\xi}_{co}$ est la moyenne arithmétique des ξ_{Kio}

dans toutes les strates et les formules (7) et (9) pour le calcul de $\sigma_{N_{kl}}^2$ sont ici:

$$\sigma_{N_{kl}}^2 = \frac{1}{12} \frac{\sum_K \sum_{t=1}^K N_{Kt} I_t P(V, N_{Kt})}{\sum_K \sum_{t=1}^K I_t P(V, N_{Kt})} \quad (13)$$

et

$$E_{N_{K1}}^2 = \frac{1}{\sum_K (N_{K1} + N_{K2})} \left[\sum_K \sum_{t=1}^K F(N_{K2t}) (N_{K2t} - q\bar{N}_0)^2 + \frac{1}{12q} \sum_K \sum_{t=1}^K F(K1t) N_{K1t}^2 \right] \quad (14)$$

où t remplace l'indice j des formules (7) et (9); ici nous avons ajouté k aux indices de chaque N ou F, et nous avons sommé pour toutes les strates. Il faut aussi noter qu'il est possible qu'il soit nécessaire d'utiliser la formule (13) pour certaines strates et la formule (14) pour les autres.

4) $\sum_K \xi_{K0}^2$ sera changé en $\frac{q^2}{p} \sum_K \xi_{K0}^2$

5) N_{ki} sera changé en qN_{ki}

6) N_k sera changé en pN_k/q

7) $\sum_K \frac{1}{N_k}$ sera changé en $\frac{q}{p^2} \sum_K \frac{1}{N_k}$

8) $\sum_K \sum_i \xi_{Kio}^2$ sera changé en $\frac{1}{pq} \sum_K \sum_i \xi_{Kio}^2$

9) A sera changé en $\frac{1}{12} \left(\frac{1}{pq} \sum_K \sum_i \xi_{Kio}^2 - \frac{16}{pq^2} \sum_K \sum_i \frac{1}{N_{Kio}} \right) = A/P1$
 A/pq

10) B sera changé en $\frac{1}{12} \left(\frac{q^2}{p} \sum_K \xi_{K0}^2 - \frac{96}{p^2} \right) \text{ car } A = p^2 \sum_K \sum_i \xi_{Kio}^2$

$$+ \sigma_{N'_{KL}}^2 \left[\xi_{..}^2 + 2p \sigma_{SKL}^2 \right]$$

11) $N_{..}$ et b/a ne seront pas changés, pour les raisons déjà données au § 2

Nous avons donc:

$$\begin{aligned} \bar{V} &= V_1 + V_2. \\ V_1 &= r(s-1) N_{..} q \sigma^2 + \frac{N_{..}(r-1) D}{q} \\ \text{avec } V_2 &= A r s^2 / p q + B r^2, \end{aligned} \quad (15)$$

où A , B et D ont été définis un peu plus haut. D'autre part:

$$C = \frac{a}{r} + \frac{b}{rs} \quad (16)$$

b) Nous choisissons des limites et des paires de valeurs convenables de p et q ,

c) Afin de calculer V_{q1} , V_{q2} et V_q , nous rendons d'abord minimum l'expression $V + \lambda c$, ce qui donne

$$\begin{aligned} \frac{2A}{pq} s_q^3 + \left(N_{..} q \sigma^2 + \frac{b}{a} \frac{A}{pq} \right) s_q^2 - \frac{b}{a} \left\{ \frac{N_{..}}{q} \left[\bar{N}_{..}^2 q^2 \sigma_{SKL}^2 \right. \right. \\ \left. \left. + \sigma_{N'_{KL}}^2 \left(\xi_{..}^2 + 2p \sigma_{SKL}^2 \right) \right] \right. \\ \left. - N_{..} q \sigma^2 + 2B r q \right\} = 0 \end{aligned} \quad (17)$$

Alors nous donnons à q diverses valeurs et nous calculons les valeurs correspondantes de s_q et de c pour chaque triplet (p, q, r) .

En interpolant, nous trouvons, pour chaque paire (p,q) , la valeur de r_q qui rend c égal à c_0 que nous avons calculé au 5.2 de la troisième partie.

Puis nous calculons les valeurs de s (d'après la formule (17) et de V_{q1} , V_{q2} et V_q V_{q1} V_{q2} correspondantes aux valeurs de r calculées dans la troisième partie.

Il est clair que la paire (p,q) correspondant au minimum de ces valeurs calculées de V est la valeur optimum de p et de q .

Application numérique: Reprenons les données de "l'enquête stratifiée TD" utilisées au § 5.2 de la troisième partie.

Nous choisissons les deux valeurs 1 et q pour p et les valeurs 1, 2, 5, 10 et 20 pour q , et nous calculons d'après la formule (17), les valeurs de s et de c pour les valeurs auxiliaires $r_q = 1, 2, 5$ et 10, séparément, pour les deux fonctions de coût (18) et (19), de la 3ème partie. Alors nous trouvons, pour chaque paire de valeurs (p,q) , par interpolation, les valeurs optimum r_q correspondant à $c_0 = 0,353a$ que nous avons déjà trouvées au § 5.2 de la 3ème partie. Nous avons porté les résultats de ces calculs dans le tableau suivant.

Valeurs optimum r_q

p		1					2				
q		1	2	5	10	20	1	2	5	10	20
c	$\frac{a}{r} \quad b \quad \textcircled{a}$	4,10	4,65	4,80	4,63	4,45	3,83	4,17	4,20	4,22	4,04
	$\frac{r}{rs}$										
c	$\frac{a}{qr} \quad b \quad \textcircled{f}$	4,10	2,86	1,47	(0,90)	(0,56)	3,83	2,26	1,18	(0,72)	(0,43)
	$\frac{qr}{rs}$										

Fonction de coût $\textcircled{a}$ (18) , $\textcircled{f}$ (19)

Pour éviter les cas pathologiques ($r_q < 1$), nous avons arrondi les valeurs $r_q < 1$ à 1; et nous avons recommencé le calcul des s_q telles que le coût soit égal à $c_0 = 0,353$, puis calculé les valeurs correspondantes de V_q .

Nous avons porté les résultats de ces calculs dans les deux tableaux suivants:

Valeurs optimum r_q , s_q et V_q quand $C = \frac{a}{r} + \frac{b}{rs}$

p	q	r_q	s_q	V_{q1} (millions)	V_{q2} (mille)			$V_q = V_{q1} + V_{q2}$
					$\frac{A r^2}{pq} q$	$B r^2$	Total	
1	1	4,10	22,4	16,0	440	49	489	16,49
	2	4,26	14,0	17,4	112	216	328	17,73
	5	4,40	12,8	42,3	45	1.440	11.485	43,78
	10	4,63	15,5	62,5	24	5.270	5.394	93,89
	20	4,45	16,6	163,3	13	19.800	19.813	183,11
2	1	3,85	29,0	24,3	345	21	366	24,67
	3	4,17	20,8	26,2	97	87	184	26,38
	5	4,20	20,6	62,5	38	551	589	63,09
	10	4,22	20,7	125,4	19	2.225	2.244	127,64
	20	4,04	21,2	240,5	10	8.150	8.160	248,66

Valeurs optimum r_q , s_q et V_q quand $C = \frac{a}{r} + \frac{b}{rs}$

p	q	r_q	s_q	V_{q1} (Millions)	V_{q2} (mille)			$V_q = V_{q1} + V_{q2}$ (Millions)
					$\frac{A}{pq} rs^2$	Br^2	Total	
1	1	4,10	22,4	16,0	440	49	489	16,49
	2	2,86	20,9	11,4	134	82	216	11,62
	5	1,47	30,0	15,2	57	135	190	15,39
	10	1	39,6	22,0	34	250	284	22,28
	20	1	33,0	36,8	12	1.000	1.012	37,81
2	1	3,83	29,0	24,3	345	21	366	24,67
	2	4,17	29,3	15,9	104	26	130	14,03
	5	4,20	45,4	17,3	52	43	95	17,39
	10	4,22	39,6	22,2	17	125	142	22,34
	20	4,02	33,0	36,8	6	500	506	37,31

On voit que (comme au § 2) notre hypothèse sur la forme de la fonction de coût (§ 2 de la 3ème partie) a une grande influence sur les résultats des calculs de r_q , s_q , et V_{q1} . Mais, dans l'un et l'autre cas, c'est-à-dire que nous adoptons la formule (18) ou la formule (19) de la 3ème partie, les résultats du calcul de V_{q2} sont presque les mêmes:

- 1) Le premier terme de V_{q2} , c'est-à-dire Ars^2/pq , croît avec q .
- 2) Le deuxième terme de V_{q2} , c'est-à-dire Br^2 , décroît quand augmente q .
- 3) Le résultat immédiat de ces deux phénomènes inverses est que V_{q2} a un minimum pour une certaine valeur de q .
- 4) Abstraction faite de la variation de q , les valeurs de V_{q2} correspondant à $p = 1$ sont plus grandes que celles correspondant à $p = 2$.
- 5) Avec la fonction de coût $c = a/r + b/rs$, la variation de V_{q2} avec q est plus rapide qu'avec la fonction de coût $c = a/q + b/rs$.

4 - Plan - Sondage à deux degrés avec stratification des PSU, tirages exhaustifs et probabilités égales.

Ici le problème est analogue à ceux étudiés au cours des précédents. Pour réduire la composante de la variance causée par la non-divisibilité des N_{ij} par les r_{ij} correspondants, nous devons augmenter la grandeur des strates et, pour cela, nous devons, soit diminuer le nombre des strates, soit diminuer le nombre des PSU, soit enfin faire les deux. La diminution du nombre des strates n'est généralement pas possible sans augmenter considérablement la valeur de la variance totale, car la variance entre les strates est très grande et c'est pour cette raison que nous stratifions les PSU, par conséquent, si nous diminuons le nombre des strates, il est possible que la variance intérieure des nouvelles strates augmente beaucoup. Nous nous contenterons donc d'étudier le cas où l'on fait varier le nombre

des PSU.

Dans le plan IV, r et les s_{ij} peuvent varier chacun indépendamment des autres, à cela près que le nombre total des éléments de la population doit rester constant. Nous devons donc rendre minimum l'expression $V + \lambda C$

La solution des équations qui expriment le minimum de $V + \lambda C$ est si compliquée dans le cas général, que nous allons seulement étudier ci-dessous les deux cas particuliers déjà examinés au § 5.3 de la 3ème partie.

4.1 Premier cas particulier.

a) Nous réunissons, en moyenne, q des PSU (qui dans notre problème ne sont pas naturelles, car elles sont elles-mêmes la réunion de plusieurs districts). Les paramètres de (74) de la 3ème partie subiront donc les modifications suivantes:

N est changé en N/q et N_i en qN_i

$$\sum N_i \eta_i^2 \text{ est changé en } \frac{1}{q} \sum q N_i \eta_i^2 = \sum N_i \eta_i^2$$

$$\sum \frac{1}{N_i} \text{ est changé en } \frac{1}{q} \sum \frac{1}{q N_i} = \frac{1}{q^2} \sum \frac{1}{N_i}$$

$$\sum \eta_i \text{ est changé en } \frac{1}{q} \sum \eta_i$$

comme au plan II,

$$\sigma_X^2 \text{ est changé en } (q\bar{N}_0)^2 \sigma_{\eta_i}^2 + \sigma_{N_i}^2 [\sigma_{\eta_i}^2 + E(\eta_i^2)] \text{ où}$$

$\sigma_{N_i}^2$ se calcule par les formules (7) et (9), tandis que $\sum \eta_i$ et X restent, exactement, les mêmes, tandis que $\sum \frac{1}{N_i}$, $\sum \frac{1}{q N_i}$, $\sum \eta_i$ et $\sigma_{\eta_i}^2$ restent approximativement les mêmes.

Les paramètres de la formule (74) de la 3ème partie deviennent donc

$$A = \beta^2 \sum_i N_i \eta_i^2 \sum_j \rho_j \tau_j^2, \quad B = \frac{3N V_m}{4 \beta^2} - \frac{V_m}{q^2} \sum_i \frac{1}{N_i} \sum_j \frac{1}{\rho_j}$$

$$E = \frac{4b}{a} \beta \frac{1}{q} \sum_i \eta_i \sum_j \tau_j = \frac{b}{a} \beta \sum_i \eta_i \sum_j \tau_j, \quad \text{et } \sigma_X^2 = (q \bar{N}_0) \sigma_{\eta_i}^2 + \sigma_{N_i}^2 \left[\sigma_{\tau_j}^2 + E(\eta_i^2) \right] \quad (18)$$

Les fonctions de variance et de coût sont donc :

$$V = V_1 + V_2$$

$$V_1 = r k \beta X - r A + \frac{N}{q} (r - 1) \sigma_X^2 \quad (19)$$

$$V_2 = r k^2 B$$

où A, B, E et σ_X^2 sont définis par la formule (18)

$$c = \frac{a}{r} \left(1 + \frac{E}{K} \right) \quad (20)$$

b) Les limites de q seront choisies au moment des applications numériques.

c) Afin de calculer r_q , s_q et V_q , nous rendons minimum $V + \lambda c$, ce qui donne

$$2 B K_q^3 + (\beta X + EB) K_q^2 - E \left(\frac{N}{q} \sigma_X^2 - A \right) = 0 \quad (21)$$

avec les mêmes paramètres que ceux donnés plus haut (système 18).

Pour chaque valeur de q, la plus grande racine de (21) donne k_q correspondant. En substituant chaque k_q dans la formule.

$$r_q = \frac{a}{c_0} \left(1 + \frac{E}{k_q} \right) \quad (22)$$

nous trouvons les valeurs correspondantes de r_q ; alors nous substituons chaque paire de valeurs (k_q, r_q) dans la formule (19) et nous trouvons les valeurs de V_q . Il est clair que la valeur de q correspondant au minimum des V_q est la valeur optimum de q .

4.2 - Deuxième cas particulier.

a) Comme dans le cas particulier précédent, notre problème est ici de trouver la valeur optimum de q . L'introduction de q change les paramètres du système (82) de la 3ème partie, de la façon suivante:

σ^2 est changé en $q\sigma^2$

N	"	"	"	N/q
σ_X^2	"	"	"	$(\bar{N}_0 q)^2 \sigma_{\xi_{i0}}^2 + \sigma_{N_{i0}}^2 [\sigma_{\xi_{i0}}^2 + E(\xi_{i0}^2)]$
$\sum_i \xi_{i2}^2$	"	"	"	$\frac{1}{q} \sum_i \xi_{i2}^2$
$\sum_i \xi_{i30}^2$	"	"	"	$\frac{1}{q} \sum_i \xi_{i30}^2$
$\sum_i \frac{1}{N_{i2}}$	"	"	"	$\frac{1}{q^2} \sum_i \frac{1}{N_{i2}}$
$\sum_i \frac{1}{N_{i3}}$	"	"	"	$\frac{1}{q^2} \sum_i \frac{1}{N_{i3}}$

tandis que les paramètres $N_{\cdot 2}$, $N_{\cdot 3}$, β , a , b , et d ne changent pas.

Les paramètres du système (82) de la 3ème partie sont modifiées comme suit:

A est changé en $N_{\cdot 2} q$ $\sigma^2 = qA$

B est changé en $N_3 q \beta \sigma^2 = qB$

$$E \quad " \quad " \quad " \quad V_m \left[\frac{1}{q} \sum_i \xi_{i20}^2 - q \sigma^2 \frac{1}{q^2} \sum_i \frac{1}{N_{02}} \right] = \frac{E}{q}$$

$$F \quad " \quad " \quad " \quad V_m \left[\frac{1}{q} \sum_i \xi_{i30}^2 - q \beta \sigma^2 \frac{1}{q^2} \sum_i \frac{1}{N_{03}} \right] = \frac{F}{q}$$

On voit que (ici aussi) les valeurs A, B et σ_X^2 qui concernent V_1 augmentent avec q , tandis que les valeurs de E et F qui concernent V_2 diminuent quand q augmente.

Les fonctions de variance et de coût deviennent:

$$V = V_1 + V_2$$

$$V_1 = r(s-1) qA + (r+1) qB + (r-1) \left\{ (\bar{N}_0 q)^2 \sigma_{\xi_{i0}}^2 + \sigma_{N_0}^2 [\sigma_{\xi_{i0}}^2 + E(\xi_{i0}^2)] \right\} \quad (23)$$

$$V_2 = \frac{r s^2 E}{q} + \frac{r t^2 F}{q}$$

où A, B, E et F ont été définis au § 5.3.2. de la 3ème partie.

$$C = \frac{q}{r} + \frac{b}{rs} + \frac{d}{rt} \quad (23)$$

b) Les limites de variation de q seront choisies à l'occasion de chaque application numérique.

c) Afin de calculer r_q , s_q , t_q , et V_q , nous rendons minimum $V + \lambda c$, ce qui donne les équations (87) et (88) de la 3ème partie où on doit multiplier A et B par q , diviser E et F par q , et remplacer σ_X^2 par l'expression

$$(\bar{N}_0 q)^2 \sigma_{\xi_{i0}}^2 + \sigma_{N_0}^2 [\sigma_{\xi_{i0}}^2 + E(\xi_{i0}^2)]$$

Comme nous l'avons expliqué au § 5.3.2. de la 3ème partie, nous pouvons résoudre les deux équations pour chaque valeur de q , et ainsi trouver s_q , et t_q . Alors, on substituera ces valeurs dans la formule

$$r_q = \frac{1}{c_0} \left[1 + \frac{b}{s_q} + \frac{d}{t_q} \right]$$

où c_0 est la valeur de c calculée au § 5.3.2. de la 3ème partie; et ainsi on trouvera r_q . Enfin on choisira comme valeur optimum, la valeur de q qui correspond au minimum de V_q .

5 - Plan V - Deux PSU (en tout) avec probabilités inégales et tirages exhaustifs.

Ainsi que nous l'avons dit à la fin du § 5.4 de la 3ème partie, ce plan ne nous procure pas de valeur optimum quand nous fixons la valeur de la variance ou du coût.

Quand nous y introduisons q , les paramètres des formules (90) et (91) de la 3ème partie changent de la façon suivante:

σ^2 est changé en $q \sigma^2$

N est changé en N/q

X_i est changé en qX_i

$\sum$ est changé en $1/q \sum$

N_i est changé en $q N_i$

$\sum X_i^2$ est changé en $q \sum X_i^2$

Ceci est tout à fait approximatif. La raison pourquoi nous avons adopté cette approximation si grossière est que le terme $\sum X_i^2$ est négligeable à côté des termes $N \cdot \sum N_i \xi_i^2$ et X^2

Les paramètres $\sum N_i \xi_i^2$, N , et A ne changent pas.

Les formules (90) et (91) de la 3ème partie deviennent

$$V(X)_q = V_{1q} + V_{2q}$$

$$V_{1q} = \sigma^2 N \cdot N(r - \frac{1}{2}) + \frac{N}{2} \sum N_i \xi_i^2 - \frac{1}{2} \sum X_i^2 - \frac{1}{2} X^2$$

$$V_{2q} = A r^2 N^2 / q^2$$

$$c = aq + b/r \quad (26)$$

qui est l'équivalent de la formule (18) de la 3ème partie. L'homologue de la formule (19) est

$$c = a + b/r$$

b) Rien à signaler.

c) Ici le résultat de l'équation $V + \lambda c$ minimum est:

$$\frac{3}{4} + \left[\frac{b}{2aq} + \frac{q N_0 \sigma^2}{2AN} \right]^2 + \frac{b}{4q^3 A q N^2} (N N_0 \sigma^2 - N_0 \sum N_i \xi_i^2 + 1 \sum X_i^2 + X^2) = 0 \quad (27)$$

La solution du système forme des équations (27) et (26) en y remplaçant c par c_0 (dont nous allons donner la méthode de calcul au § suivant) nous procure les valeurs optimum de q et de r_q ; et en substituant ces valeurs dans la formule (25), nous trouvons les valeurs V_{q1} , V_{q2} et V_q minimum.

6 - PLAN VI - Stratification, puis tirage de 1 PSU de chaque strate avec probabilités inégales.

Comme on est ici obligé de soumettre à l'enquête, dans chaque strate, une seule PSU, on est toujours intéressé à changer la taille des strates afin de pouvoir modifier le nombre des PSU soumises à l'enquête. Essayons de résoudre ce problème.

a) Réunissons, en moyenne, q strates en une seule.

Les paramètres des formules (94) et (95) de la 3ème partie subissent les changements suivants:

N_k est changé en $q N_k$

N " " " N/q

$$\sum_K N_k N_{k_0} \quad " \quad q \sum_K N_k N_{k_0}$$

$$\frac{1}{N} \quad " \quad " \quad q \frac{1}{N}.$$

$$\sum_K N_{k_0} \sum_L N_{kl} \bar{X}_{kl_0}^2 \quad " \quad q \sum_K N_{k_0} \sum_L N_{kl} \bar{X}_{kl_0}^2$$

$$\sum_K X_K^2 \quad " \quad " \quad q \sum_K X_K^2$$

Cette transformation ici est raisonnable, car dans ce paragraphe nous utilisons $\sum_K X_K^2$ toujours sous la forme du second terme de

$$\sum_K N_{k_0} \sum_L N_{kl} \bar{X}_{kl_0}^2 - \sum_K X_K^2$$

qui est la variance entre les PSU à l'intérieur des strates (et nous avons traité ainsi les variances depuis le début).

$$\sum_K \frac{1}{N_{K_0}} \sum_L N_{KL} \sum_{K_0}^2 \text{ est changé en } \frac{1}{q} \sum_K \frac{1}{N_{K_0}} \sum_L N_{KL} \sum_{K_0}^2$$

$$\sum_K \frac{N_K}{N_{K_0}} \quad " \quad " \quad " \quad \frac{1}{q} \sum_K \frac{N_K}{N_{K_0}}$$

$$A \quad " \quad " \quad " \quad A/q$$

$$B = \sum_K N_{K_0} \sum_L N_{KL} \sum_{K_0}^2 - \sum_K X_K^2 - \sigma^2 \sum_K N_K N_{K_0} \text{ est changé en } qB$$

Les formules (94) et (95) de la 3ème partie sont remplacées

par

$$\begin{cases} V = V_1 + V_2 \\ V_1 = \bar{N} \cdot rN \cdot \sigma^2 q + qB \\ V_2 = \bar{N}^2 \cdot Ar^2 q \end{cases} \quad (28)$$

et

$$c = a/q + b/rq \quad (29)$$

Il faut noter que, en adoptant la formule (29), nous avons renoncé à notre troisième hypothèse du § 2 de la 3ème partie : nous avons préféré ici la fonction (19) à la fonction (18) (et nous n'avons donc ici de calculs que pour une seule fonction de coût), car l'augmentation de q ne change pas les frais généraux d'une strate.

b) Avec la méthode de calcul que nous allons donner, nous n'avons pas besoin d'indiquer les limites pour q .

c) Afin de trouver la valeur optimum de q , ou bien nous transformons les paramètres de la formule (97) de la 3ème partie en y introduisant les transformations mentionnées ci-dessus, ou

bien nous rendons directement minimum l'expression $V + \lambda c$ en y substituant les formules (27) et (28), pour les variations de q et de r . Le résultat des deux procédés est le même, à savoir:

$$r_q^3 + r_q^2 \left[\frac{b}{2a} + \frac{N \dots 5^2}{2a\bar{N}} \right] - \frac{b}{2a\bar{N}} \frac{B}{A} = 0 \quad (30)$$

équation du 3ème degré en r indépendante de q .

Alors nous calculons q tel que le coût de l'enquête soit égal à c_0 , c'est-à-dire:

$$q = \frac{1}{c_0} \left(a + \frac{b}{r_q} \right) \quad (31)$$

Dans la 3ème partie nous n'avons pas été en mesure de calculer une valeur pour c_0 . Donc, ici nous calculons une valeur qui remplace c_0 . Nous calculons c_0 pour une valeur fixe V_0 que nous donnons à V_1 , en rendant minimum l'expression

$V_1 + \lambda c$ pour les variations de q et r , où V_1 provient de la formule (28) et c de la formule (29). Les résultats de ces calculs sont:

$$r_0 = \sqrt{\frac{b}{a\bar{N}.N \dots 5^2} + \frac{B}{A}} \quad (32)$$

$$q_0 = \frac{V_0}{\bar{N}_0 r_0 N \dots 5^2 + B} \quad (33)$$

$$c_0 = \frac{1}{q_0} \left(a + \frac{b}{r_0} \right) \quad (34)$$

Ce sera la valeur $\hat{\alpha}_0$ calculée par la formule (34) que nous substituerons dans (31).

Application numérique. - Reprenons les données de "l'enquête stratifiée TD" utilisées au § 5.5. de la 3ème partie. Les résultats des calculs sont:

1) En utilisant les formules (32), (33) et (34), nous avons calculé les différentes valeurs optimum pour la valeur fixe de V_1 égale à $V_0 = 16.000.000$ (pour un coefficient de variation 5 %).

Ces valeurs sont:

$$r_0 = 14,6 \quad , \quad q_0 = 1/33,3 \quad \text{et} \quad c = 56,1 \text{ a}$$

C'est-à-dire que nous devons augmenter le nombre des strates de 5 (structure présente) à $5 \times 33,3 = 166$ (structure optimum si nous ne tenons pas compte des erreurs d'arrondissement). Dans ce cas, chaque strate nouvelle comprendra, en moyenne, $1.282/166 = 7,72$ villages.

2) En utilisant c_0 calculé ci-dessus et les formules (31) (28) et (29), nous trouvons:

$r_q = 13,6$, q = valeur optimum de q quand nous tenons compte des erreurs d'arrondissement (soit $1/32,4$), $V_{q1} = 16.070.000$, $V_{q2} = 345.000$ et $V_q = 16.415.000$

Donc, quand nous tenons compte des erreurs d'arrondissement, mais voulons que le coût reste le même ($c_0 = 56,1a$), la structure de la population doit être telle que:

Le nombre des strates nouvelles soit $5 \times 32,4 = 162$, et chaque strate nouvelle renferme $1.281/162 = 7,8$ villages (en moyenne).

La comparaison entre ces deux séries de valeurs numériques montre que (1) pratiquement les erreurs d'arrondissement sont négligeables et (2) si cependant on tient compte des erreurs d'arrondissement, on doit diminuer très légèrement le nombre des strates et augmenter la fraction de sondage.

7 - Plan VII.- Sondage à deux degrés, avec probabilités inégales et tirages bernoulliens.

Au point de vue structure optimum, ce problème est analogue à celui du plan II.

a) Quand nous introduisons q dans les paramètres des formules (98) et (99) de la 3ème partie, ils se modifient exactement comme ceux du plan V (§ 5), avec cette différence apparente qu'ici la fonction de coût, d'après la définition donnée au § 5.6 de la 3ème partie, est:

$$c = aqn + Nbnq/r \quad (35)$$

La formule (98) de la 3ème partie est changée en

$$V = V_1 + V_2 \quad (36)$$

$$V_1 = \frac{1}{n} q \left(N \cdot \left(r - \frac{N}{q} \right) + \frac{N \cdot}{n} \sum_{i=1}^N x_i^2 - \frac{X^2}{n} \right)$$

$$V_2 = \frac{VM}{n} \frac{X^2}{2} + \frac{Vm}{N \cdot n} r^2 \sum_{i=1}^N x_i^2 + \frac{r^2}{N \cdot} \frac{Vm}{2} q \sum_{i=1}^N x_i^2$$

b) Rien à signaler

c) En rendant minimum l'expression $V + \lambda c$, nous avons

$$\begin{aligned} & \frac{3}{2q} \left[\frac{2n_q V_m}{N_0} \sum_i N_i \xi_{i0}^2 + \frac{2n_q^2 V_m q}{N_0^2} \sum_c K_c^2 \right] \\ & + \frac{2}{2q} \left[n_q N_0 q \xi^2 + n_q \frac{b}{a} \frac{N V_m}{q N_0} \sum_i N_i \xi_{i0}^2 \right. \\ & \quad \left. + 2n_q^2 \frac{b}{a} \frac{N}{N_0^2} V_m \sum_c X_c^2 \right] \\ & - \left[n_q \frac{b}{a} \frac{N N_0}{q} \sum_i N_i \xi_{i0}^2 + \frac{2b}{a} \frac{N}{q} V_m X^2 \right. \\ & \quad \left. - n_q \frac{b}{a} \frac{N^2}{q} N_0 \xi^2 - n_q \frac{b}{a} \frac{N}{q} X^2 \right] = 0 \quad (37) \end{aligned}$$

Pour utiliser la formule (37), nous procéderions comme suit:

- 1) Nous choisirions des valeurs convenables de q entre des limites légitimes.
- 2) Nous choisirions aussi des valeurs convenables de n_q et, pour chaque paire de valeurs (q, n_q) choisies, nous calculerions n_q d'après la formule (37) et puis c d'après la formule (35).
- 3) Pour chaque valeur de q nous calculerions par interpolation, la valeur de n_q correspondant à la valeur $c = c_0$ calculée au § 5.6 de la 3ème partie.
- 4) Pour chaque n_q ainsi calculé, nous calculerions r_q par la formule (37) et V_q par la formule (36).

La valeur de q correspondant au minimum de V_q donnera la taille optimum des PSU.

8- Influence du changement de la structure de la population sur V_2 contribution des erreurs d'arrondissement à la variance.

Comme nous l'avons dit déjà pour la 3^{ème} partie (§ 6), une comparaison mathématique entre les résultats de la 4^{ème} partie, en ce qui concerne les valeurs optimum, est impossible pour nous.

Pour tirer des résultats généraux des études de cette partie, nous pouvons:

a) étudier les résultats des applications numériques, ce qui a été déjà fait assez complètement, à la fin des paragraphes correspondants; et nous n'allons pas donc y revenir; et

b) comparer les expressions de V_2 que nous avons obtenues après avoir changé la structure de la population.

Nous allons faire ci-dessous cette dernière comparaison.

Comme dans le § 9 de la deuxième partie, nous examinons le cas abstrait des populations où V_1 est identiquement nul.

En:

1) admettant les approximations du § mentionné;

2) Considérant les fractions de sondage comme fixes;

3) posant: q = nombre moyen des PSU naturelles qu'on réunit pour en construire une nouvelle, et p = nombre moyen des strates naturelles qu'on réunit pour en constituer une nouvelle; et

4) posant $V'_2 = V_2$, après avoir changé la structure de la population (c'est-à-dire après avoir introduit les paramètres q et p);-

Nous avons trouvé des expressions assez simples et comparables pour les V'_2 des 7 plans étudiés (le plan I, dans l'introduction et les plans II à VII, dans la 4ième partie)

Plan	Expression de V'_2	Plan	Expression de V'_2
I	V_2/p	V	V_2/q
II	V_2/q	VI	$p V_2$
III	$(K/q + L q^2)/p$	VII	$K'/q + L' q^2$
IV	V_2 / q		

Nota 1- Dans ce tableau, $K, L, K',$ et L' sont des constantes renfermant les fractions de sondages (considérées comme fixes) et les paramètres de la population ; et V_2 est l'expression de la contribution des erreurs d'arrondissement à la variance avant changement de la structure de la population (avant d'introduire p et q).

Nota 2- Contrairement aux autres plans, nous avons pour le plan VII introduit dans la formule (36) (§ 7), pour l'échantillon des PSU, l'effectif n au lieu de la fraction de sondage (par exemple $1/s_1$), et au lieu de la fraction de sondage des éléments dans les PSU (par exemple $1/s_2$), nous avons introduit r , qui est lui-même égal à nR , où $1/R$ est la fraction totale de sondage de la population (voir § 9 de la 2ième partie). En tenant compte de ce fait, c'est-à-dire en posant

$R = s_1 \quad s_2, n = N/s_1$ et $r = n R$, nous avons trouvé, pour le V_2' de la formule (36), une expression comparable à celles des autres plans; et c'est elle qui est portée dans le tableau.

Résultats.

1) Pour les plans I, II, IV et V où nous avons tenu compte d'une seule fraction de sondage comme source des erreurs d'arrondissement, les résultats sont les mêmes (ce qui est conforme au bon sens; car quand on augmente la taille q des PSU, ou la taille p des strates à l'intérieur desquelles on doit sonder, V_2' décroît, naturellement, avec $1/q$ ou $1/p$).

2) Pour le plan VI, on constate un phénomène qui semble être en contradiction avec les précédents, tandis que, en réalité il est seulement d'une nature différente. Ici nous avons changé la taille des strates dont les effectifs d'échantillon sont toujours entiers ($=1$). D'autre part, nous n'avons pas changé la taille des PSU, à l'intérieur desquelles les effectifs de sondage sont non-entiers. La variance intérieure des PSU reste donc la même. Mais comme, dans le cas du plan VI, le résultat d'une seule PSU doit représenter le total d'une strate, la variance de l'estimation est proportionnelle au nombre total des PSU se trouvant dans cette strate; et par conséquent, V_2' , de même que V_1 (voir la formule 36), doit être proportionnelle à p .

3) Les V_2' des plans III et VII ont la même forme (ne différant que par présence et absence de p). La raison de leur ressemblance est que pour ces deux plans (contrairement à ce qui se passe pour les autres plans) nous avons tenu compte simulta-
nément

de deux fractions de sondage comme sources des erreurs d'arrondissement. Les termes comprenant K et K' qui décroissent avec $1/q$, sont les contributions à la variance des erreurs d'arrondissement des effectifs de l'échantillon dans chaque PSU; ces termes doivent, naturellement, décroître quand la taille des PSU augmente (et c'est le phénomène déjà constaté pour les plans I, II, IV et V). Mais les termes contenant L et L' proviennent de l'arrondissement des nombres de PSU qui doivent être soumises à l'enquête; et ils doivent naturellement, croître avec q ; car plus grand sera q et plus petit sera le nombre des PSU.

On trouverait aussi les mêmes sortes de termes, proportionnel à q^2 , pour les plans II et IV, si on n'y avait pas fait l'abstraction de la contribution des erreurs d'arrondissement du nombre des PSU échantillon.

La différence formelle entre les V'_2 des plans III et VII (c'est-à-dire la présence et l'absence de p) est le résultat du changement de la taille des strates dans le plan III. Ici, comme dans le plan I et contrairement au plan VI, l'augmentation de la taille des strates a fait décroître, avec $1/p$, la valeur de V'_2 ; puisque, dans le cas des plans I et III, on voulait soumettre à l'enquête dans chaque strate un nombre non-entier de PSU ou d'éléments.

Conclusions.

1- Quelle que soit la méthode (non-biaisée) employée pour arrondir les effectifs théoriques de l'échantillon à l'un ou l'autre des entiers les plus proches, la variance V_m de l'effectif adopté

pour chaque strate ou unité première de sondage (PSU)

- a) dépend seulement de la méthode d'arrondissement utilisée,
- b) est comprise entre les 2 limites

$$\frac{1}{12} - \left[\sqrt{\frac{1}{42}} - \frac{1}{4} \right] \leq V_m \leq \frac{7}{12} + \left[\sqrt{\frac{1}{12}} - \frac{1}{4} \right]$$

2- Nous avons étudié 5 méthodes d'arrondissement dont les V_m sont comprises entre les limites

$$\frac{1}{12} \leq V_m \leq \frac{7}{12}$$

3- Si nous arrondissons les effectifs théoriques d'échantillon d'après l'une de ces méthodes (non-biaisées), et utilisons cependant le même estimateur sans biais qu'on a obtenu sans tenir compte des erreurs d'arrondissement, nous obtenons une estimation encore sans biais, mais telle que sa variance, V , renferme deux termes, V_1 V_2 , où

V_1 est la même expression qu'on trouverait si on n'avait pas tenu compte des erreurs d'arrondissement, et

V_2 est un terme complémentaire dépendant de V_m .

Nous avons vérifié, en étudiant 7 plans différents de sondage, que V_2 est croissante avec V_m , ce qui nous permet d'adopter la méthode dont V_m est la plus petite, comme étant la meilleure.

4- Plus le nombre de découpages de la population est grand (pour une stratification ou un degré de sondage p), et plus V_2 sera grande.

5- Plus le nombre d'unités d'échantillon est petit, et plus V_2 est grande.

6- Lorsqu'on veut tenir compte des erreurs d'arrondissement, on doit utiliser de nouvelles formules de "l'optimum"; et dans ce cas on a:

6a) Quand on calcule les valeurs optimum des fractions de sondage sans tenir compte des erreurs d'arrondissement, la variance sera peut-être de 2 à 6% supérieure à la valeur désirée V_0 de la variance introduite dans les calculs " de l'optimum".

6b) Quand on tient compte des erreurs d'arrondissement dans les calculs des valeurs optimum des fractions de sondage, théoriquement on obtient un certain gain de variance; mais pratiquement ce gain est beaucoup trop faible pour compenser la contribution des erreurs d'arrondissement à la variance.

6c) Pour être sûr que la variance V ne dépasse pas la valeur désirée V_0 , on devrait, dans les calculs " de l'optimum " remplacer V_0 par une valeur plus petite et telle que la variance définitive soit justement égale à V_0 .

7- Comme la taille des strates et des PSU à une certaine influence, d'une part sur V_1 et d'autre part sur V_m et par conséquent sur V_2 , on peut toujours trouver une structure optimum de la population en recherchant la taille optimum des strates et des PSU.

Nous avons étudié ce problème sur 7 plans différents de sondage et nous avons vérifié que la structure de la population a des effets différents sur V_1 et V_2 , de sorte que la structure

optimum de la population existe bien toujours; mais sa détermination nécessite pour chaque plan la résolution d'un système d'équations convenable.

8- Quand on réunit plusieurs PSU naturelles pour en construire une nouvelle:

8a) La contribution (à la variance) des erreurs d'arrondissement des effectifs théoriques de sondage à l'intérieur des PSU, décroît quand la taille des PSU augmente.

8b) La contribution (à la variance) des erreurs d'arrondissement du nombre théorique des PSU, ainsi construites, croît avec la taille des PSU.

9- Quand on change la taille des strates à l'intérieur desquelles on devrait soumettre à l'enquête un nombre théorique non-entier de PSU ou d'éléments, la contribution (à la variance) des erreurs d'arrondissement de ces nombres décroît quand la taille des strates augmente.

10- Quand on change la taille des strates à l'intérieur desquelles on se propose de soumettre à l'enquête un nombre fixe des PSU, la contribution (à la variance) des erreurs d'arrondissement, de même que celle des autres erreurs, croît avec le nombre des PSU dans chaque strate.

Références.

- 1- The Experimental sample survey of the Households of Teheran, June 20- June 24, 1954. Public Statistics Cooperative Organization, Teheran.
- 2- Mahalanobis, P.C. (1940), A sample survey of the Acreage under Jute in Bengal, Sankhya, 4, 511-30
- 3- Smith H. Fairfield (1938). An Empirical Law Describing Heterogeneity in the Yield of Agricultural Crops, Jour. Agr. Sci. 28, 1-23.
- 4- P. Thionet, Recherche d'une fonction de coût applicable aux enquêtes de l'E.N.S.I.I., Revue de l'Institut International de statistique, 1954:1/3, P. 33-47.
- 5- Cochran, W.G. Sample Survey Technique. North Carolina State College, Institute of Statistics. Mimeo Series n°7; 1950.
- 6- Sampling Theory of Survey With Applications. By Pandurang V. Sukhatme, Food & Agricultural Organization of the United Nations.
- 7- Yates, F. 1949. Sampling Methods for Censuses and Surveys. Hafner Publishing Company, New York.
- 8- Bulletin de Statistique, n°18, Ministère de l'Intérieur; Administration Générale des Statistiques et de l'Etat-Civil de l'Iran, Janvier-mars 1953.
- 9- Khajehnouri, A.G, Sample survey of Water Used for Agriculture in Iran, Raleigh 1952. Thèse non-publiée présentée à la faculté de "North Carolina State College" pour le grade de M.Sc.

10- Gumbel, E.J. (1934) Les valeurs Extrêmes des Distributions Statistiques. Annales de l'Institut Henri Poincaré, 5, 115.

11- Fisher, R.A. & Tippett, L.H. (1934) Limiting Forms of the Frequency Distribution of the Largest Member of A Sample. Proc. Camb. Phil; Soc. 24, 180.

RESUME DE LA DEUXIEME THESE

Présentée

sous la direction de M. GEORGES DARMOIS

de l'Académie des Sciences,

Professeur de CALCUL DES PROBABILITES

et Physique Mathématique.

DEVANT LA FACULTE DES SCIENCES

DE L'UNIVERSITE DE PARIS

Pour l'obtention du titre de Docteur de l'Université

(mention Sciences)

par

ABBAS GHOLI KHADJENOURI

=====

L'orthogonalité et ses applications à la
statistique mathématique.

Au cours de nos années d'études, nous nous sommes aperçus que les notions d'orthogonalité auxquelles il est fait allusion dans différents ouvrages et qui jouent un rôle important dans les études de la statistique mathématique, au même titre que le calcul matriciel et le calcul des probabilités, etc..., n'ont été rassemblées dans aucun ouvrage. Afin de faciliter son enseignement et son application systématique, nous avons rassemblé ces notions, et pour en faire une théorie indépendante nous avons été amenés à poser quelques définitions et notations nouvelles.

Cet exposé comprend onze chapitres dont nous donnons les résumés ci-dessous.

Résumé du chapitre I.-

Au cours du chapitre I, nous avons donné les définitions de l'espace d'échantillon E_n , du point d'échantillon F_n et du vecteur d'échantillon ; voici cette dernière:

Définition du vecteur d'échantillon.

Désignons par $\bar{x}$ la moyenne arithmétique des mesures

$$X = (x_1, x_2, \dots, x_n)$$

des éléments - échantillon et par A le point de E_n dont les coordonnées sont $(\bar{x}, \bar{x}, \dots, \bar{x})$. Désignons par P le point d'échantillon, dont les coordonnées sont x . Nous définissons le vecteur d'échantillon par AP

Considérons plusieurs "caractéristiques" $X, Y, \dots$ et posons:

$$\xi = [\xi_1, \xi_2, \dots, \xi_n]$$

$$\eta = [\eta_1, \eta_2, \dots, \eta_n]$$

etc,...

avec $\xi_i = x_i - \bar{x} \quad ; \quad \eta_i = y_i - \bar{y} \quad ; \quad \text{etc} \dots$

1) ξ, η etc... sont des vecteurs dont nous dirons qu'ils ont une certaine liberté.

2) Le nombre de degrés de liberté du vecteur d'échantillon sera, par définition, égal à $n-1$.

3) On posera:

$$\xi' \xi = n l_{xx}$$

$$\eta' \eta = n l_{yy}$$

$$\xi' \eta = n l_{11}$$

où l_{20} , l_{02} et l_{11} sont les moments d'ordre 2 d'échantillon.

En désignant par θ la mesure de l'angle formé par les vecteurs ξ et η ; et par r le coefficient de corrélation entre X et Y, nous avons:

$$\cos \theta = \xi' \eta / [\xi' \xi \eta' \eta]^{\frac{1}{2}} = r$$

d'où la définition de l'orthogonalité de deux "ensembles" d'éléments-échantillon en statistique:

Définition: Deux "ensembles" d'éléments-échantillon sont dits orthogonaux l'un à l'autre si

$$\xi' \eta = 0$$

Nous étendons la définition donnée ci-dessus à m "ensembles" $x_1 = (x_{11}, x_{12}, \dots, x_{1m}), \dots, x_m = (x_{m1}, x_{m2}, \dots, x_{mm})$ et disons que les m vecteurs $x_1, x_2, \dots$ et x_m constituent un système d'ensembles orthogonaux si

$$x_i' x_j = a_{ij} \delta_{ij} \quad (1)$$

(où δ_{ij} est le delta de Kroneker).

Enfin nous avons dit que:

La condition nécessaire et suffisante pour que toutes les corrélations entre m "caractéristiques" étudiées soient nulles est que leurs vecteurs d'échantillon soient orthogonaux dans E_n .

Alors nous avons donné la définition classique des fonctions orthogonales:

$$g(x) = [g_1(x), g_2(x), \dots]$$

par la relation:

$$\int_S g_i(x) g_j(x) dF(x) = a_{ij} \delta_{ij} \quad (2)$$

()

(où S est le champ de variation de x); et nous avons dit que (1) est un cas particulier du système(2).

Résumé des chapitres 2 et 3.

Au cours des chapitres 2 et 3, nous avons donné (avec démonstration) les théorèmes algébriques qui nous ramènent à connaître les transformations et les matrices orthogonales qui ont beaucoup d'applications en statistique mathématique.

D'abord nous avons donné comme suit une définition des matrices orthogonales plus générale que la matrice orthogonale des transformations orthogonales.

Définition: Deux matrices A et B sont dites orthogonales, au sens général, si:

$$A'B = (a_{ij} \delta_{ij})$$

ce qui entraîne:

$$A'B = B'A$$

(Nous avons démontré au chapitre 3 que la matrice orthogonale des transformations orthogonales est un cas particulier des matrices ainsi définies).

Les théorèmes classiques que nous avons donnés (avec leurs démonstrations) au chapitre 2 sont:

Théorème 1.- Une forme quadratique $Q(x,x)$ dont le rang est r , peut toujours être transformée par une transformation linéaire et non-singulière (que nous désignerons ci-dessous par TLNS), à une forme canonique

$$Q(x,x) = \sum_{i=1}^r \alpha_i X_i^2$$

Théorème 2.- Si l'on transforme k fois une même forme quadratique $Q(x,x)$ de rang r par des TLNS différentes en des formes

$$\sum_{i=1}^r \alpha_{ij} X_i^2 \quad [j=1, 2, \dots, k]$$

le nombre des α positifs, dans les k formes ainsi obtenues, est le même; il en est de même du nombre des α négatifs. Il résulte de là que, en désignant par P le nombre des α positifs et par N le nombre des α négatifs, la "signature" $s = P - N$ de la forme $Q(x,x)$ reste constante quand nous changeons la TLNS.

Théorème 3.- Pour chaque forme définie positive $P(x,x)$ dont la matrice est d'ordre m ; il y a une TLNS de module un, qui la transforme à la forme canonique:

$$\sum_{i=1}^m c_i X_i^2$$

où tous les coefficients C_i sont positifs.

Théorème 4.- Soit A une matrice, d'ordre $m \times n$ ($m \leq n$) et de rang m , dont les éléments sont réels (matrice des observations:

m caractéristiques de n éléments d'échantillon). La matrice $A'A$ (c'est-à-dire le produit de la matrice des moments par n) a les propriétés suivantes:

- a) Elle est symétrique.
- b) Elle est définie positive.
- c) Ses racines caractéristiques sont positives.

Théorème 5. - Soit $Q(x,x) = x'Qx$ une forme quadratique et $P(x,x) = x'Px$ une forme définie positive, toutes deux d'ordre m. Nous pouvons toujours trouver une TLNS qui transforme, en même temps, ces deux formes respectivement en

$$\sum_{i=1}^m \lambda_i X_i^2 \quad \text{et} \quad \sum_{i=1}^m X_i^2$$

où λ_i sont les racines de l'équation $\lambda'(|Q - \lambda P| = 0)$ déduite des matrices Q et P.

Les théorèmes 1 à 5 nous servent de bases pour étudier et comprendre les matrices et les transformations orthogonales;

Les théorèmes 1 et 2 nous donnent des renseignements sur le rang et la "signature" des formes quadratiques.

Le théorème 3 nous enseigne ceux des formes définies positives.

Le théorème 4 nous sert à comprendre le rôle essentiel de la transformation orthogonale en statistique.

Enfin le théorème 5 nous aide à définir la transformation orthogonale. Il nous dit que nous pouvons toujours trouver une TLNS qui mette simultanément la forme quadratique $Q(x,x)$ sous sa

forme canonique $\sum_{i=1}^m \lambda_i X_i^2$ et la forme définie positive $P(x,x)$ sous sa forme canonique $\sum_{i=1}^m X_i^2$ où les λ_i sont les racines de l'équation λ des deux formes.

Au chapitre 3, nous avons étudié le sous-ensemble de l'ensemble des TLNS mentionnées ci-dessus, qui mettent la forme quadratique $Q(x,x)$ sous sa forme canonique $\sum \lambda_i X_i^2$ les λ_i étant, cette fois-ci, les racines caractéristiques de la forme $Q(x,x)$. En d'autres termes, nous nous débarrassons de l'interférence de la forme définie positive $P(x,x)$.

Nous avons démontré que le sous-ensemble recherché est l'ensemble des transformations orthogonales classiques.

Définition : Si la matrice C d'une transformation linéaire $x = Cy$ est telle qu'après la transformation, on ait:

$$x'x = y'y$$

la transformation est appelée une transformation orthogonale, et sa matrice C une matrice orthogonale.

Au cours du chapitre 3, nous avons donné les théorèmes classiques concernant la transformation orthogonale, que nous avons résumés comme suit:

$$1) CC' = C'C = I$$

$$2) C' = C^{-1}$$

3) Si λ est l'une des racines caractéristiques de C , $1/\lambda$ l'est aussi.

4) Pour chaque matrice carrée A d'ordre m , on peut toujours trouver une matrice orthogonale C telle que la matrice $C'AC$ soit une matrice diagonale ayant comme éléments $\lambda_1, \lambda_2, \dots, \lambda_m$ les racines caractéristiques de la matrice A .

5) Le nombre des racines caractéristiques d'une matrice est égal à son rang.

6) Soit $Q(x, x)$ une forme quadratique de m variables x ; supposons que $m-r$ de ces variables dépendent linéairement des autres r variables. Nous pouvons toujours transformer cette forme, par une transformation orthogonale, en une forme canonique $\sum_{i=1}^r \lambda_i X_i^2$ où les nouvelles variables X (au nombre de r) sont linéairement indépendantes les unes des autres.

C'est cette transformation importante qui est souvent utilisée dans les calculs statistiques et est, en effet, à l'origine de l'importance de la transformation orthogonale en statistique mathématique.

7) La "signature" d'une forme quadratique est égale au nombre de ses racines caractéristiques positives, moins celui de ses racines caractéristiques négatives.

8) Chaque forme définie positive $P(x, x)$ d'ordre m a (m) racines caractéristiques toutes positives.

9) Les transformations utilisées au cours des théorèmes du chapitre 2 sont, à des facteurs près, des transformations orthogonales.

10) En utilisant la notion de transformation orthogonale, nous avons donné la représentation géométrique des formes qua-

quadratiques et définies positives, laquelle sera utilisée dans les chapitres suivants.

11) Nous avons aussi donné les conditions pour construire des matrices orthogonales (théorèmes de Cayley).

12) Les modules des racines caractéristiques de toute matrice orthogonale sont réels.

Résumé du chapitre 4.

Ici nous avons seulement donné quelques théorèmes et définitions classiques de statistique, qui ne concernent pas l'orthogonalité, afin de pouvoir les utiliser au cours des chapitres suivants.

Ces théorèmes et définitions sont: Degré de liberté-Distribution des fréquences - Transformation de variables - Moments - Matrice des moments du deuxième ordre - Fonction caractéristique Rang de la distribution - Transformation linéaire des variables Un théorème sur la distribution en limite - Inégalité de Bienaïme Tchébitcheff - Ellipsoïde de concentration - Distributions : normale, t , X^2 et F .

Résumé du chapitre 5.

Au cours du chapitre 5 nous avons donné:

1) Théorème de Cochran - Soit $x = (x_1, x_2, \dots, x_n)$ n variables indépendantes. Si on partage la forme $\sum_{i=1}^n x_i^2$ (dont le rang est, manifestement, n) en k formes quadratiques non-négatives $Q_1, Q_2, \dots, Q_k$ dont les rangs sont respectivement

$r_1, r_2, \dots, r_k$ tels que $r_1 + r_2 + \dots + r_k = n$, de sorte qu'on a:

$$\sum_{i=1}^n x_i^2 = Q_1 + Q_2 + \dots + Q_k$$

on peut toujours trouver une transformation orthogonals qui

transforme, en même temps, $\sum_{i=1}^n x_i^2$ en $\sum_{i=1}^n y_i^2$

et chaque Q_j en :

$$\sum_{i=r_1+r_2+\dots+r_{j-1}+1}^{r_1+r_2+\dots+r_j} y_i^2$$

2) Lemme de Fisher - Soit $x = [x_1, x_2, \dots, x_n]$ n variables aléatoires normales indépendantes. Considérons p paramètres

$\eta = [\eta_1, \eta_2, \dots, \eta_p]$ calculés en utilisant les variables x et les relations

$$\eta = C_1 x$$

où C_1 est une matrice d'ordre $p \times n$ dont les éléments c_{ij} sont tels que les relations d'orthogonalité existent entre ses lignes. Si nous transformons les x en $n-p$ autres variables

$$\xi = [\xi_1, \xi_2, \dots, \xi_{n-p}]$$

par la formule de transformation

$$\xi = C_2 x$$

telle que la matrice :

$$C = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

soit orthogonale, on établit alors que:

a) Les variables $y = [\eta | \bar{s}]$ sont elles aussi normales et indépendantes les unes des autres, avec la même variance que les x .

b) Le rang de la forme quadratique

$$Q = x'x - \eta'\eta$$

est égal à $n-p$.

Dans le même chapitre en même temps que nous avons donné toutes les explications et démonstrations nécessaires pour ces deux théorèmes (qui sont les bases théoriques de l'analyse de variance), nous avons illustré, par des exemples, la mécanique de l'application de l'orthogonalité.

3) Nous avons donné aussi les parties des démonstrations des formules d'intégration suivantes qui concernent l'orthogonalité:

$$\begin{aligned} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} e^{it'x - (1/2)x'Ax} dx_1 \dots dx_m &= \frac{(2\pi)^{\frac{m}{2}}}{\sqrt{|A|}} e^{(1/2)t'A^{-1}t} \\ \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} e^{-(1/2)x'Ax} dx_1 \dots dx_m &= \frac{(2\pi)^{\frac{m}{2}}}{\sqrt{|A|}} \\ \int_{x'Ax < \rho^2} dx_1 \dots dx_m &= \frac{\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2} + 1)} \frac{\rho^m}{\sqrt{|A|}} = V \end{aligned}$$

$$\int_{\substack{\dots\dots\dots \\ \chi' A \chi < P^2}} \chi_i \chi_j dx_1 \dots dx_m = \frac{f^2 V}{m+2} \frac{A_{ij}}{|A|}$$

Résumé du chapitre 6.

Ici nous avons retrouvé, en utilisant l'orthogonalité, les lois de distribution des variables aléatoires dérivant de la variable normale. Les variables étudiées sont : la variable normale elle-même et les variables de Student, χ^2 , F , $\bar{x}$, $s^2 = \sum (x - \bar{x})^2 / n$, $\bar{x}_i$ et f_{ij}

Dans tous les cas, et spécialement dans les cas où le nombre de degrés de liberté de χ^2 est $(n-k-1)$, nous nous sommes contentes de donner les calculs concernant l'orthogonalité.

Résumé du chapitre 7.

Au cours du chapitre 7, nous avons donné les calculs de l'analyse de variance d'une façon plus générale que les calculs classiques. Nous avons étudié le cas d'un découpage multiple de la variable suivant le modèle statistique:

$$x_{i_1 \dots i_{k+1}} = \mu + \alpha_{i_1} + \dots + \alpha_{i_1 \dots i_k} + \epsilon_{i_1 \dots i_{k+1}} \quad (1)$$

Nous avons adopté les notations suivantes:

$$N(t, s) = \sum_{t_1} \dots \sum_{t_s} N_{i_1 \dots i_s}$$

$$X(t) = \sum_{i_1} \cdots \sum_{i_{k+1}} x_{i_1 \dots i_{k+1}}$$

$$\bar{x}(t) = X(t)/N(t, k)$$

$$X(t)^2 = X^2(t)/N(t, k)$$

$$X_{q0}(t)^2 = \sum_{i_1} \cdots \sum_{i_t} X(t)^2$$

$$X_0(t)^2 = X_{q0}(t)^2 \Big|_{q=0}$$

$$\xi(t) = \bar{x}(t) - \bar{x}(t-1)$$

Alors nous avons donné au moyen de huit lemmes, les règles d'algèbre de ces notations.

Enfin nous avons démontré le théorème suivant:

L'ensemble $\xi = [\xi(1), \xi(2), \dots, \xi(k+1)]$ (où $\xi(t)$ est celui qui a été défini plus haut) est un ensemble de fonctions orthogonales; c'est-à-dire que:

$$\sum_{i_1} \cdots \sum_{i_{k+1}} \xi(t) \xi(t') = \delta_{tt'} SS_t$$

où $SS_t = X_0(t)^2 - X_0(t-1)^2$ est la variance entre les éléments de découpage d'ordre t à l'intérieur des éléments d'ordre $t-1$.

En utilisant ce théorème avec le théorème de Cochran et le lemme de Fisher et en désignant le degré de liberté correspondant à SS_t par dl_t , nous avons démontré que:

$$\sum_{i_1} \cdots \sum_{i_{k+1}} x_{i_1 \dots i_{k+1}}^2 = SS_0 + SS_1 + \dots + SS_{k+1}$$

$$dl_t = N(0, t-1) - N(0, t-2)$$

et

$$N(0, k) = dl_0 + dl_1 + \dots + dl_{k+1}$$

Ces résultats classiques nous ont permis d'admettre que, sous l'hypothèse nulle (les α du modèle (1) étant nuls et les variables e normales ayant une moyenne nulle et une variance σ^2) la loi de la distribution du ratio

$$dl_t, SS_t / dl_t SS_t = \chi^2_{dft} / \chi^2_{dft'}$$

est une "loi de F", ce qui nous permet d'utiliser les tables de F (ou z) pour tester l'hypothèse nulle.

Nous avons aussi donné la formule générale du calcul des espérances mathématiques des carrés moyennes

$$MS_t = SS_t / dl_t$$

Cette formule est:

$$E(MS_t) = \sigma^2 + \sum_{r=1}^K K_{tr} \sigma_r^2$$

avec

$$K_{tr} = \frac{1}{dl_t} \left[\begin{array}{l} N(0, K) - \sum_{i_1} \dots \sum_{i_t} \frac{N^2(t, k)}{N(t-1, k)} \quad \text{si } t = r \\ \sum_{i_1} \dots \sum_{i_r} N^2(r, k) \left[\frac{1}{N(t, k)} - \frac{1}{N(t-1, k)} \right] \quad \text{si } t < r \\ 0 \quad \text{si } t > r \end{array} \right]$$

et avec :

$$\sigma_r^2 = E(x_{i_1}^2 \dots x_{i_r}^2)$$

Résumé du chapitre 8.

D'abord nous avons donné les définitions classiques pour l'additivité, l'effet et l'interaction des variables.

Puis nous avons donné les lois classiques pour tester les différentes hypothèses concernant la relation linéaire entre les variables. En posant:

$$y = X\beta + e$$

où

$$y = [y_1, y_2, \dots, y_n] = \text{valeurs observées de la variable dépendante,}$$

$$e = [e_1, e_2, \dots, e_n] = \text{erreurs}$$

$$\beta = [\beta_1, \beta_2, \dots, \beta_m] = \text{coefficients de régression,}$$

et

$$X = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nm} \end{pmatrix} = \text{valeurs observées des variables indépendantes;}$$

et en utilisant les notions d'orthogonalité, nous avons donné les formules classiques pour tester les hypothèses:

$$1) \beta = 0$$

$$2) \beta = \beta_0$$

$$3) \text{ une partie des } \beta = \gamma = \gamma_0$$

Enfin nous sommes arrivés à la relation classique d'orthogonalité:

$$x'_{\alpha} x_{\alpha'} = S_{\alpha\alpha'} \delta_{\alpha\alpha'}$$

comme condition nécessaire et suffisante pour que le test de l'hypothèse sur chaque partie des β , c'est-à-dire l'hypothèse

$$\beta^{(\alpha)} = \beta_0^{(\alpha)} \quad \text{soit indépendante de nos hypothèses}$$

ou informations à propos des autres parties des β

$$\text{Ici } \beta^{(1)} = [\beta_1^{(1)}, \dots, \beta_{m_1}^{(1)}], \dots, \beta^{(k)} = [\beta_1^{(k)}, \dots, \beta_{m_k}^{(k)}];$$

sont les k partitions de β ; les m_{α} sont tels que

$\sum_{\alpha=1}^k m_{\alpha} = m = \text{nombre des variables}$; les x_{α} sont les k partitions de x telles que x

les $S_{\alpha\alpha'}$ sont des matrices des nombres et $\delta_{\alpha\alpha'}$ sont les symboles de Kroneker.

Résumé du chapitre 9.

En partant de la dernière formule et en utilisant, directement, les notions d'orthogonalité, nous avons étudié les plans classiques d'expérimentation: simple, en bloc, carré latin et carrés orthogonaux d'ordre plus élevé.

Résumé du chapitre 10.

Au cours de ce chapitre nous avons discuté et donné les lois et les formules classiques concernant l'utilisation des polynômes orthogonaux pour la régression.

Résumé du chapitre 11.

Dans ce chapitre nous avons étudié différentes fonctions orthogonales dont la plupart ont été utilisées en statistique. Pour chaque fonction nous avons donné la définition et quelques propriétés mathématiques ne concernant pas l'orthogonalité mais nécessaires pour démontrer la propriété de l'orthogonalité; et enfin nous avons démontré la relation d'orthogonalité de chaque fonction.

Les questions étudiées sont:

Cas général- La série de Fourier - Calcul des coefficients d'une série de fonctions orthogonales par la méthode des moindres carrés (dont les calculs du 10ème chapitre est un cas particulier) - Utilisation de la série de Fourier pour calculer les formules de corrections de Sheppard - Les polynômes de Tchébitcheff- Hermite - La série de Gram-Charlier - Les fonctions de Ressel de la première sorte - Les polynômes de Legendre - Les fonctions orthogonales de l'analyse harmonique.





